山东师范大学

本 科 生 毕 业 设 计

题　　目：基于 Python 的二手房数据分析预测系统

姓　　名：常子儒

学　　号：201911990102

专　　业：计算机科学与技术（公费师范生）

指导教师：魏艺

学院（部）：信息科学与工程学院

2023 年　5　月　20　日

独 创 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得_____（注：如没有其他需要特别声明的，本栏可空）或其他教育机构的学位或证书使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　导师签字：

学位论文版权使用授权书

本学位论文作者完全了解 **学校** 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **学校** 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签字：

签字日期：20　年　月　日　　签字日期：20　年　月

目录

基于 python 的二手房数据分析预测系统

常子儒

（山东师范大学信息科学与工程学院 2019 级计师本 1901 班）

摘要：本文以潍坊市二手房交易市场为研究对象开发了基于 python 的二手房数据分析预测系统，旨在为房地产行业提供数据、图像支持和决策依据。本系统通过搭建 Flask 框架，利用 requests、lxml 等第三方库从链家等二手房交易网站上爬取潍坊市二手房信息并存储，将获得的数据信息进行清理并作预处理，然后利用 numpy、pandas、matplotlib 等数据分析库对预处理完成的数据作可视化来了解数据的特征与分布，最后利用 Scikit-learn 机器学习库建立多元线性回归模型与 K 近邻模型并训练优化，利用交叉验证来评估模型性能。本文建立的二手房数据分析预测系统提供数据查询、数据图表展示及房价预测等功能，使用户能深入挖掘二手房数据信息并帮助其做出决策。

关键词：二手房交易；Flask 框架；数据分析；数据可视化；线性回归

Analysis and prediction system of second-hand housing data based on python

Ziru Chang

(School of Information Science and Engineering, Shandong Normal University)

Abstract: This article develops a Python-based second-hand housing data analysis and prediction system for the Weifang second-hand housing market, aiming to provide data, image support, and decision-making basis for the real estate industry. The system uses the Flask framework to crawl second-hand housing information from websites such as Lianjia in Weifang through third-party libraries such as requests and lxml, and then stores the obtained data information. The data is cleaned and preprocessed, and then visualized using data analysis libraries such as NumPy, Pandas, and Matplotlib to understand the characteristics and distribution of the data. Finally, the Scikit-learn machine learning library is used to establish a multiple linear regression model and K-nearest neighbor model, which are trained and optimized using cross-validation to evaluate model performance. The second-hand housing data analysis and prediction system established in this article provides functions such as data query, data chart display, and house price prediction, enabling users to explore second-hand housing data information in depth and help them make decisions.

Key words: Second-hand house; Flask framework; Data Analysis; Data Visualization; Linear Regression

1. 绪论

1.1 课题研究背景及意义

随着社会经济迅速发展，二手房市场迅猛发展，交易量居高不下。二手房既是住房，也可以被用以理财投资。中国城镇化建设加速，我国城镇率早已超过 60%，各式各样的居民楼拔地而起，这也意味着有越来越多的农村居民前往城镇，在城镇安家工作。从农村走出的年轻人大部分都选择了在城市买房，迫切需要更多的住房。2023 年已有近 30 个城市调整了首套房贷利率。优化限购、降低首付、补贴税费、提高贷款额度也逐渐成为调控政策优化的“标准配置”，众多政策的到来反映了各地房地产市场正在活跃起来，二手房市场也已出现回暖的迹象，2023 年前两个月，全国二手房市场也保持着一定的成交热度。

国内多家研究所发表了二手房交易的统计数据，根据克而瑞研究中心数据，全国 11 个重点城市的二手房成交较去年春节月的成交增长 29%，出现回暖迹象。从成交数据来看，1 月份青岛和东莞二手房成交量均实现同比正增长，成都、苏州、佛山、杭州、深圳等 9 个城市成交量超过 2022 年春节月 2 月份；据贝壳研究院统计，1 月贝壳 50 城二手房价格指数环比上涨 0.2%，这是自 2021 年 8 月连续 17 个月环比下跌以来首次止跌。50 城中约六成城市二手房价格 1 月止跌，覆盖京津冀、长三角、粤港澳大湾区、成渝等重点城市群。

度过了前几年楼房交易市场的低迷，房地产行业及二手房交易正在逐渐恢复，加之利好政策鼓励住房消费来刺激楼市交易，必然会使得人们购房意愿逐步增强，人们对于二手房房产价格及相关信息评估的需求也会随之增大，建立一个针对二手房数据的分析预测系统是很有帮助的。

因为近年来潍坊经济发展迅速，基础设施建设不断完善，城镇化成果显著，二手房交易也非常频繁，对数据获取比较便利，因此本文选择潍坊市作为研究对象，对其二手房交易市场进行数据分析及房价预测。

1.2 国内外研究现状

1.2.1 国外研究现状

国外在很早的时候就有专门从事二手房交易的职业--二手房中介与房屋经纪人，相关企业起步早，软件技术发达，有完善的二手房信息管理软件。

1993 年初 Matthew Gray's Wandered 在麻省理工学院开发出第一个网络爬虫，开启了数据采集自动化的进程，为二手房数据收集提供了便利。2013 年 David Gray 运用谱分析的方法确定了房价的动态变化，研究是否可以在爱尔兰城市的二手住宅房地产市场上找到房价连锁效应的证据[1]，2019 年 Jon Stobart 借鉴北安普敦郡 1761-1836 年拍卖记录，以确定二手纺织品数量和性质变化，以及涨价与估价方式，并揭示了国家房屋拍卖是二手商品再循环中的一个关键机构[2]，2019 年 Raul-Tomas Mora-Garcia 等人分析和量化了阿利坎特市场上二手房的要价与表征它们的属性之间的关系，结果表明，阿利坎特市场的价格分割，北部沿海地区的价格高于南部和内陆[3]，2020 年 Christian A.L. Hilber, Olivier Schöni 调查了自然舒适度高的地方对于二手房投资者政治限制的影响，利用实验“瑞士第二家园计划”得到研究结果，季节性旅游地点对二手房市场的发展做出限制[4]。2021 年 Koktashev Vladislav 等学者对二手房市场价格取决因素进行了比较，利用机器学习的非参数方法，构建预测模型，层次聚类等方法，实现了公寓成本预测的高精度，揭示和描述了二手住房对象价格形成的特殊性[5]。

1.2.2 国内研究现状

国内二手房数据分析研究开始时间与国外相比相对较晚。

2000 年刘明吉，王秀峰和黄亚楼针对当年随数据库与人工智能发展起来的新兴学科——数据挖掘做出了解释，将对数据处理的研究放于数据挖掘研究工作的重点[6]。随后数据挖掘中数据预处理研究也逐渐增加：2014 年以董倩为代表的学者对北京、上海、天津、重庆等 16 个大中城市的二手房和新房价格进行了研究。基于中国最大的搜索引擎百度搜索指数，他们进行了数据处理和分析。他们使用支持向量机（SVM）和随机森林对收集的数据进行回归预测，并对其进行进一步拟合[8]，在数据挖掘领域，2015 年原继东和王志海提出了针对时间序列的非数据适应性表示方法、数据适应性表

示方法和基于模型的表示方法，并详细介绍了基于时域相似性、形状相似性和变化相似性的分类算法，此外，他们还展望了未来的研究方向[9]。随着计算机技术的不断发展，大数据时代带来的数据组织模式与类型多样化，以及数据质量参差不齐等问题，给数据感知表达、理解和计算带来了巨大挑战。在 2018 年，以孔钦为代表的学者对数据预处理的主要任务进行了分析，总结了对“脏数据”的几种常用处理方法，并阐释了在数据清洗、集成、变换和归约过程中的常用算法[11]。

数据采集方面，爬虫技术成为了一种非常强大的数据抓取工具。在 2020 年，徐志和金伟通过对网络爬虫原理的阐述，利用 Python 爬虫技术对网页数据进行了抓取[13]，同年钟机灵也开发了基于 Python 网络爬虫技术的数据采集系统，实现了主题数据的自动采集。该系统采用了 urllib、Beautiful Soup、threading 等库，设计和开发了包含数据爬取、异常处理、robots 协议管理和多线程管理等模块的系统模型框架[14]；2022 年，姬正骁运用 Python 爬虫工具对链家网上武汉市各行政区在售二手房数据进行了采集，随后，他进行了数据清洗，并使用了 Matplotlib 和 Pyecharts 库进行了数据可视化分析[17]，同年，洪丽华和黄琼慧探讨了如何利用 Python、爬虫技术和网页爬虫等方面的知识，帮助用户搜索和整理相关数据[18]。

在进行数据采集和预处理后，将数据以更直观的方式呈现出来非常重要。2006 年王薇指出视觉信息图表将成为未来数据可视化的主流，她通过分析信息时代的视觉需求以及阐述视觉信息图表的概念，强调了信息时代视觉图表设计的价值[7]；指出视觉信息图表将成为未来数据可视化的主流。她通过分析信息时代的视觉需求以及阐述视觉信息图表的概念，强调了信息时代视觉图表设计的价值[19][20]，他们的研究有助于更好地理解如何利用数据可视化技术来有效传达和解释数据，同时也为数据分析人员提供了更丰富的工具和资源。

数据预测是数据挖掘领域的一项深层次研究。多元线性回归算法是数据挖掘中有效的一种算法，也是机器学习中基础性的回归算法。2016 年，冷建飞、高旭和朱嘉平将多元统计分析作为基础和前提，验证了相关结果改变对于多元线性回归方程整体的影响，并通过实例对模型进行了检验，提高了准确度和效率，使回归结果得到最大程度上的优化[10]。相较于传统的单项预测，组合预测能够有效地集成各项预测方法的信

息，在预测实践中有广泛的应用。2019 年，王自成、朱家明和陈华友改进了组合预测模型方法，提出了逐步回归筛选的回归组合预测模型。这种方法改变了自变量进入方程的方式，有效地筛选出对实验结果不显著的变量，并利用人口数据进行实例分析，证明了筛选后的回归组合预测方法有更高的预测精度[12]。2020 年，以崔明明为代表的学者采用组合集成预测的方法对房地产市场的变化方向和水平进行了预测[15]，2021 年，李函谕、魏嘉银和卢友军针对深圳市二手房市场房价预测问题，结合相关的八个特征变量，利用随机森林模型训练了房价预测模型，并得出了二手房市场信息变动的结论[16]。这些研究有助于更好地理解如何利用数据预测技术来分析和预测不同领域的现象，并提供了有益的思路和方法。

1.3 论文主要研究内容

本课题以潍坊市作为研究对象，对潍坊市二手房交易市场进行了调查，通过爬虫从链家等网站获取潍坊市挂牌出售的二手房交易信息，并开发二手房数据分析预测系统，本小节将简述系统开发目标及要完成的主要工作内容。

本课题需要完成的目标：对爬取的二手房数据做分析预测，然后将搭建系统实现上述功能并进行扩展。在开发系统前需先进行系统需求分析，阐明系统需要完成的具体功能，罗列出详细的待实现功能，随后将开发任务分为两部分，第一部分为数据处理与数据分析，这部分包括数据采集和预处理、特征工程和数据分析、建模和预测；第二部分为构建系统，本文选择基于 HTML、CSS 和 JavaScript 前端开发语言搭建 web 网站，利用 bootstrap 前端开发和 layui 模块化前端 UI 框架等完成界面布局设计，增强网页内容的美观性与可读性。两部分任务互不分离，根据系统需求分析明确数据分析的范围与内容，同时根据数据分析所得到的结果确定系统模块并完成相应的功能。

2. 需求分析

对系统做需求分析是整个项目开发的第一步，本文研究的问题为设计开发二手房数据分析预测系统，因而在开始设计系统之前先对潍坊市二手房市场做了一定的考察，确定系统开发需要完成的目标功能以及要如何实现系统功能，这样才能确保后面的系

统设计与系统实现能顺利进行。

确定本系统的目标受众，这部分人大多为二手房购房者或房屋中介等，在需求分析时应紧密地与用户联系，根据用户的具体需求描述去实现相应的功能，对于用户反映的系统缺陷也要找到合适的方法积极完善改进。该系统的主要任务是帮助用户查询二手房信息，给予用户可视化图表展示，并能根据用户不同要求给出二手房房价预测信息。

2.1 功能性需求

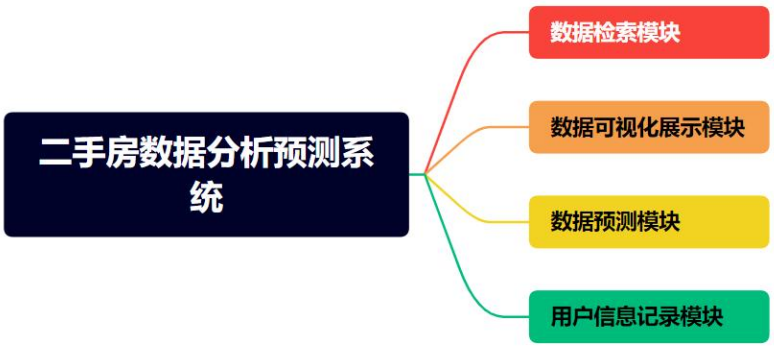


图 2-1　功能需求模块分析图

图 2-1 使用思维导图来对二手房数据分析预测系统的功能性需求进行描述，可以更加清晰明了地说明系统具有的功能模块。

通过对主流二手房信息网站一系列的调查分析，对用户需求研究后，并结合潍坊市二手房市场实际情况，本文设计的系统需要包含以下功能：

2.1.1 数据查询模块

将获取的二手房数据信息做一定的整理，完成二手房信息的录入，修改与删除功能，例如获取房源的简要介绍信息、房屋总价、房屋单价、房屋面积、房屋户型、房屋朝向等信息，按照一定的逻辑编排成特殊格式，导入数据库中，可以通过对数据库存储的信息做修改以达到更新内容的功能。用户则可以在客户端或浏览器操作图形化界面，按照意愿输入自己想要搜索的内容，从而系统展示特定的搜索内容给用户。

2.1.2 数据可视化展示模块

大多数用户可能对于具体且繁多的数据信息不太敏感，在数据查询时总是无法找到某些数据的特征或数据特征不够明显，因而通过以图像的形式展示数据信息各个角度的特征以及对数据特征做横向与纵向对比，能很好地帮助部分用户掌握数据特征及其之间存在的规则联系。

2.1.3 数据预测模块

二手房房价是备受购房者关注的一个要素，绝大多数购房者都对房价可能的走势比较关心，本文对二手房市场大多购房者的需求调研，确定系统需要完成对特定面积、特定户型、特定朝向、特定地区的二手房信息进行相应预测的功能，按照用户筛选的房源信息，为用户建立算法模型并根据模型训练结果给出房价参考值。

2.1.4 个人信息记录模块

作为一个数据分析系统，保留用户的登录与注册信息并对用户下次登录做识别是基本的功能性需要。当用户在注册时可以提供的电子邮箱，系统可以通过给邮箱发送验证码信息，来确保电子邮箱可用，并完成接下来的注册工作。当用户在注册时输入内容不符合格式，系统会对用户的不合规输入做提示，直到按照要求完成注册。当用户在输入信息并登录时，系统会将输入内容与数据库等原有存储的注册信息做比对，若相符合，则登录通过，若不符合，则登陆失败，同时系统可以对错误内容做反馈来帮助用户完成登录。

2.2 非功能性需求

图 2-2 非功能性需求分析图

图 2-2 使用思维导图来对系统的非功能性需求进行描述，可以更清楚地说明各个需求的作用。

除了上述提到的功能性需求之外，如何保证系统的正常运行、数据信息的完整性以及满足用户体验感等非功能性需求也是检验一个系统完成度高低的重要标准，以下几点为该系统可能需要满足的非功能性需求：

2.2.1 更新性

二手房数据信息纷繁错杂并且日新月异，从二手房交易信息发布到二手房交易达成，这个过程不是漫长的，而是在一个可计量的时间范围内，所以它是具有时效性的，因而如何确保已经交易二手房的数据信息不再显示与新的二手房信息发布这两个工作能顺利进行是一个值得思考的需求问题。二手房信息的不断变化，这就要求系统应该具有良好的拓展性与更新性，以满足系统中二手房信息及用户数据的不定时更新，以及增加新的系统功能。

2.2.2 安全性

身处信息时代的今天，信息泄露可能普遍存在，一个系统如果没有安全的运行环境或维护手段可能是非常危险的，安全性是保证系统正常运行的保证。假如用户信息泄露，则用户身份可能会被顶替，甚至不法分子会利用泄露的用户名、密码等信息去其他网站撞库，暴力破解其他关联系统。因而系统需要验证用户登陆信息的合法性，以及用户在系统存储的身份记录信息安全无泄漏，同时应该保证用户数据不丢失，及时、按时、实时做备份，当数据信息误删时能够第一时间恢复；还要保证数据信息不被篡改，防止无关人士故意修改数据信息，因而需要对数据存储的工具做权限限制，保证只有系统管理人员才能修改数据。

2.2.3 系统性能

保证系统流畅运行需要系统具有较高的性能，例如系统载入时间缩短与系统数据提交返回不受限，这可以从提高服务器配置，优化系统代码等方向来开展工作。

2.2.4 界面需求

一个 UI 设计美观简洁，功能部分清晰明确的系统界面是受用户喜欢的很重要的因

素，系统开发完成后它的使用者是用户而不是开发者，因此在开发系统前做好调查与广泛参考成功的系统页面的设计，确定自己独特新颖的系统界面来吸引用户，同时完善交互功能，便于操作。

3. 系统设计

系统设计是前期对用户需求分析的一个在功能与非功能性上的详细设计过程，它包括系统从前期规划到开始搭建框架、设计如何完成各功能模块以及最后该如何部署上线等问题，本章将对系统总体工作实现流程及具体功能进行介绍。

3.1 总体设计流程

下图 3-1 按照工作流程描述了数据处理及系统搭建的任务：

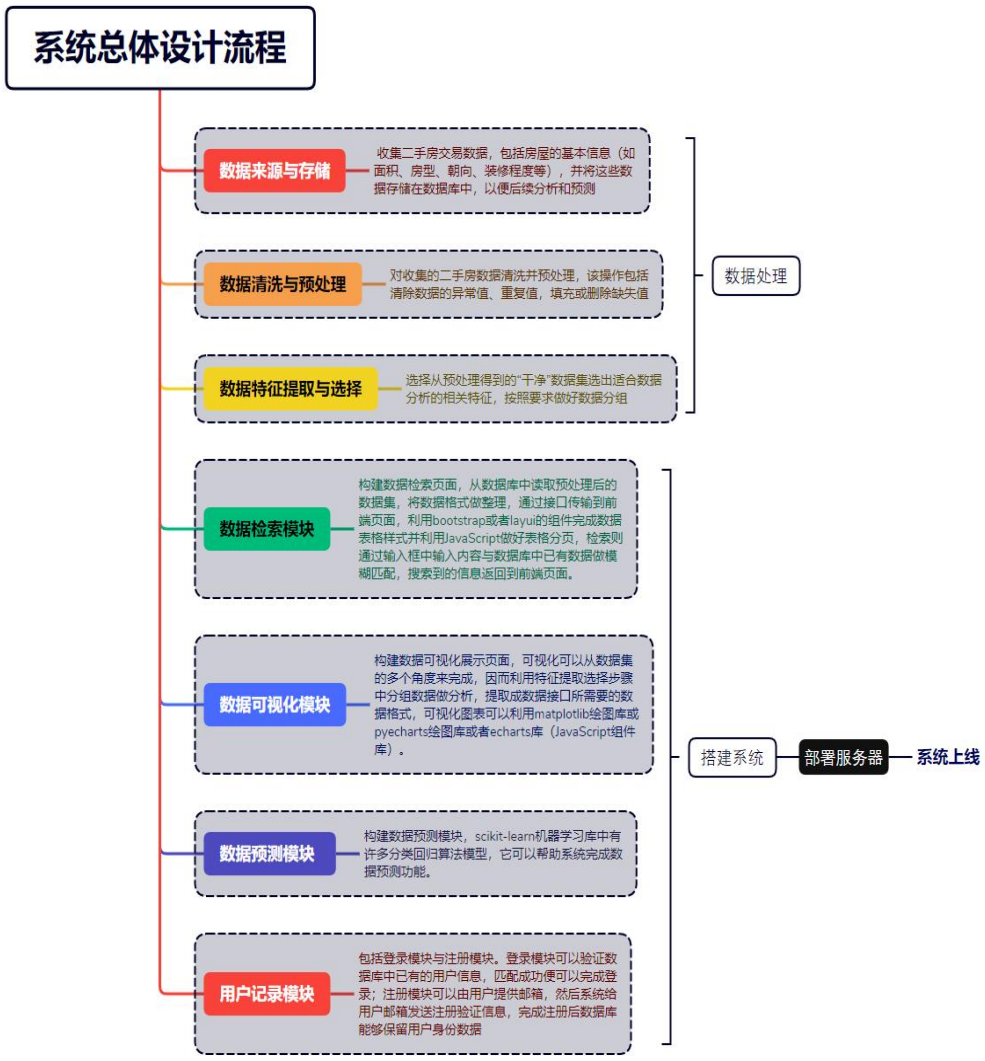


图 3-1　总体设计流程图

系统总体设计包括两个部分，第一部分为前期的数据处理工作，该工作包括搜集二手房数据，确定数据来源，通过爬虫将数据从二手房信息网站爬取下来并保存到数据库中，按照命名原则对数据库库名、表名、字段名等命名，数据库保存后，对数据做预处理，预处理的主要对象为重复值、异常值和缺失值等，按照数据特征做好分组，将不需要的特征数据删除，方便后续进行数据可视化和预测工作；第二部分为搭建系统网站，以需求分析确定的数据检索模块、数据可视化模块、数据预测模块和用户信息记录模块的基础，利用前后端相关技术实现各模块功能，在本地整体搭建完系统后，最后选择服务器，将项目系统部署到服务器中完成系统上线，实现人人都可以访问。为了更直观的了解系统的结构组成以便开发者的利用。

3.2 系统功能设计

本平台主要以 python 作为开发语言，利用后端 Flask 框架，搭建本地网站，从而形成以网站为基础的数据分析系统，将系统程序划分为几种功能模块并以网站页面的形式展现，各模块需要满足高内聚、低耦合的设计原则。

3.3 界面设计

合理的页面布局与美观的界面效果能更好地吸引用户并且使用户操作更加方便，在经过对相关二手房信息网站的调查研究，确定网站页面的基本布局与相关组件。

在正式编写前端代码构建网站页面之前，先利用电脑自带的画图软件对需要建设的网页效果做了一个简单的绘制工作，确定网页中大概需要具有什么内容。

如图 3-2 所示，设计的页面是用户进入网站最先看到的首页，该部分内容应该包括对网站的简要介绍，网站的最新消息，二手房有关的政策、新闻等动态，以及对网站功能模块的指向性。

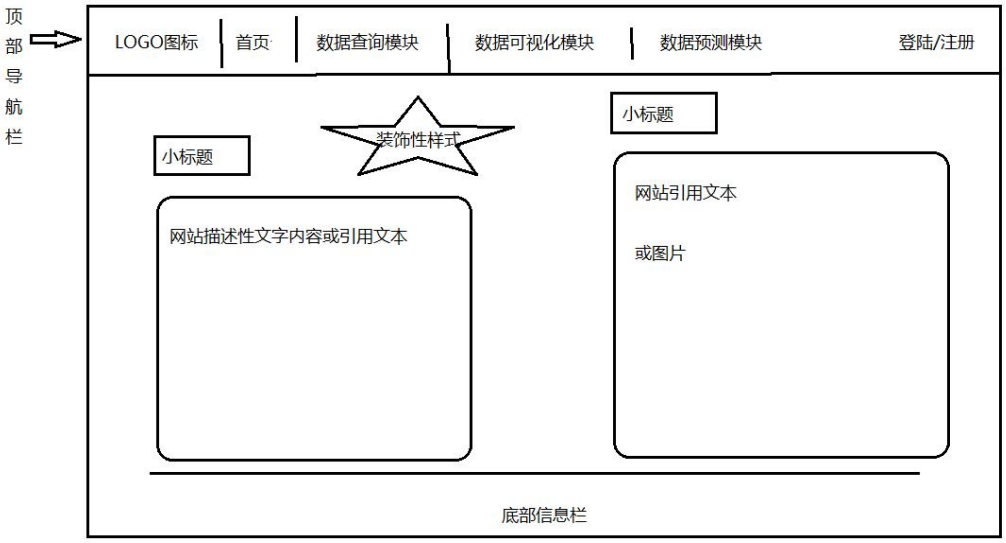


图 3-2　网站首页设计图

如图 3-3 所示，设计内容为数据查询功能的页面设计图，界面首先要有一个标题用于介绍本页面的主要内容，除此之外，在该页面中主要有两部分：用户的输入和查询信息的展示，这是在用户需求分析中的一个功能模块，当初始数据信息过多时，如果数据全部展示，可能会导致数据加载十分缓慢因而必要时可以再对数据表格做分页功能以确保每个页面展示数据的数量不会过多也不会过少。

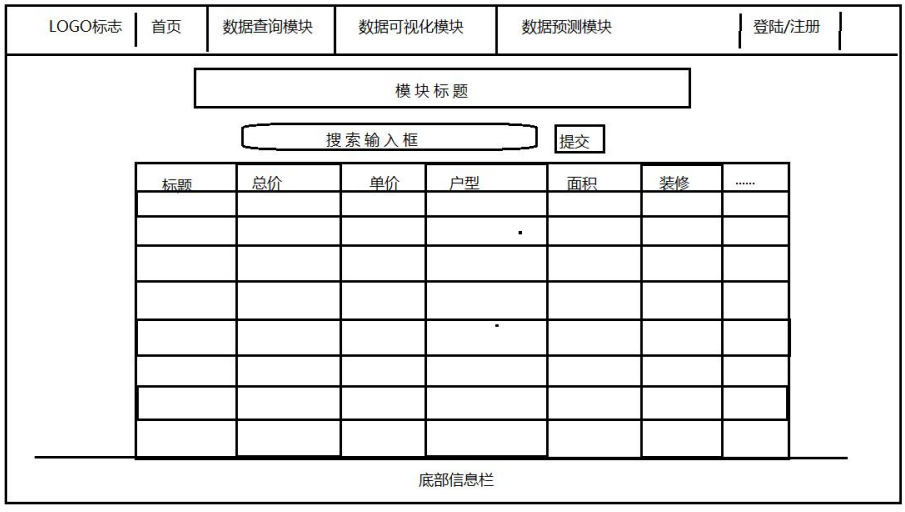


图 3-3　数据查询模块设计图

如图 3-4 所示，设计内容为数据可视化模块的页面设计图，在该模块中需要利用可视化工具 matplotlib 或 pyecharts 又或者 Echarts 等从多个角度、多个数据特征展示各要素之间的变化关系与数量关系。

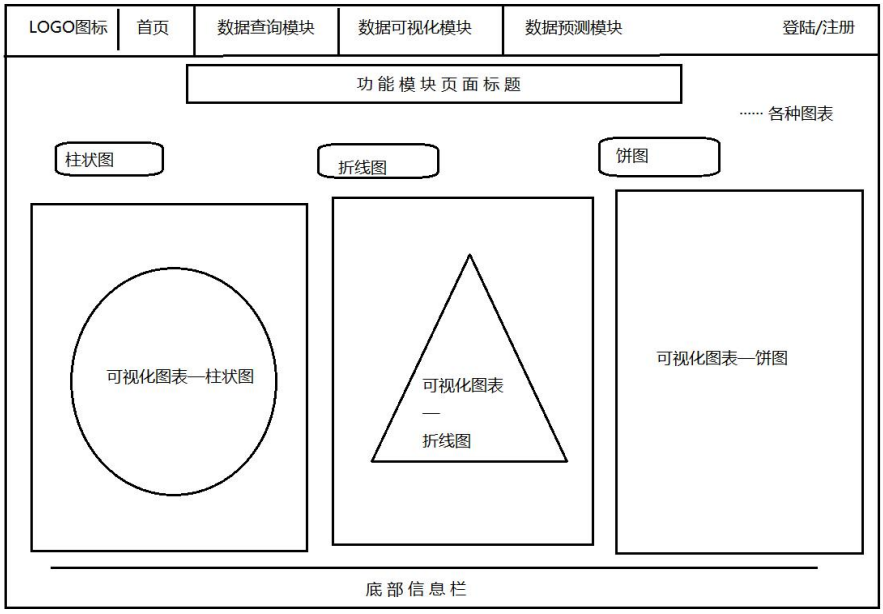


图 3-4　数据可视化模块设计图

如图 3-5 所示，设计内容为数据预测模块的页面设计图，界面由标题栏与功能内容组成，预测功能具体版块由机器学习训练出的模型图片、用户输入的条件信息版块和模型根据输入数据信息输出的预测信息版块三部分组成，使界面结构看起来条理清晰、模块化程度高。

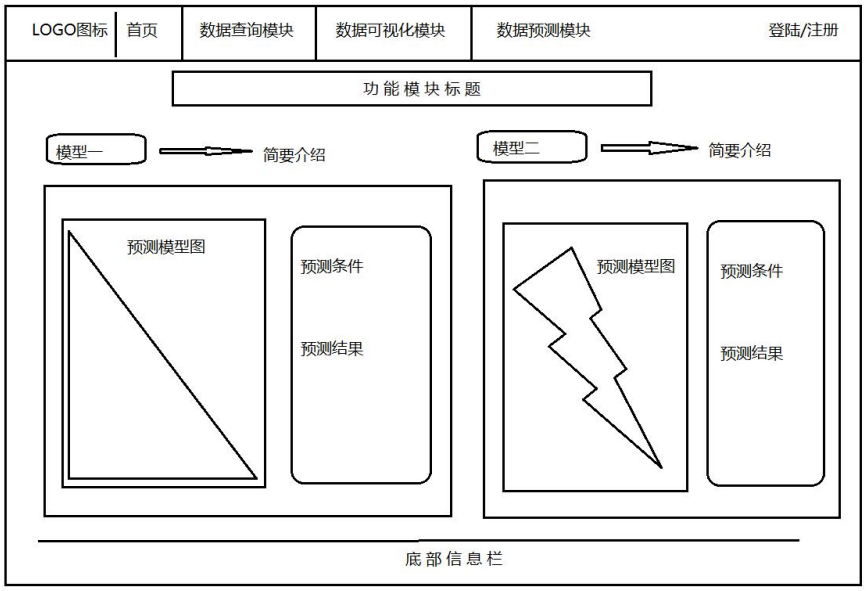


图 3-5　数据预测模块设计图

3.4　数据库设计

二手房数据信息在确定来源并获取之后，便需要将数据做存储工作。存储数据的

形式有好多种，例如用 CSV 文件存储、数据库存储、Excel 表格文件存储或者是 JSON 文件存储等等，如此多的存储方式中较为常用的便是数据库存储，原因显而易见，比如数据库可以将数据集中存储在一个地方以方便做数据管理与维护工作；与其他存储形式相比，数据库有较高的安全性，数据可以及时备份与加密；此外强大的数据查询与数据分析也是数据库一个非常重要的功能。

上述指的是数据库有巨大的优势，那数据库从无到有应该有一个完整的设计流程，因而通过分析得出这个过程应为从概念设计到逻辑设计再到最后的物理设计，如图 3-6 所示。

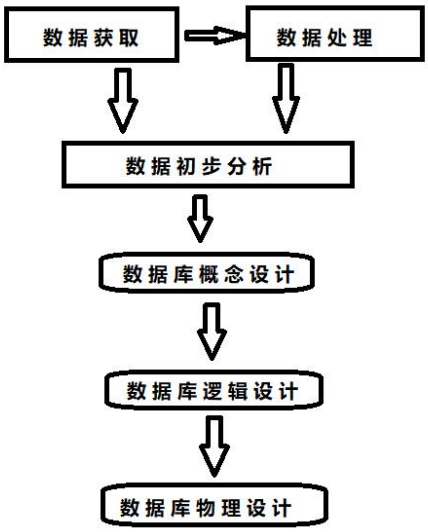


图 3-6　数据库设计流程图

（1）概念设计：根据需求分析结果、二手房数据网站及二手房交易市场的二手房信息，可以确定数据库中哪些数据需要、哪些不需要。数据库中应包含两部分数据，一部分是二手房信息，这部分应该有二手房的简介、总房价、房屋单价、房屋面积、户型、房屋朝向、装修情况、电梯有无、梯户比例、房屋性质等等属性，另一部分是用户信息，这部分数据是指用户在注册时系统给用户发送的验证码信息，这些数据会存到数据库中用于邮箱验证，另外用户在系统注册完成后，数据库会记录个人身份信息用于比对用户身份。

（2）逻辑设计：首先需要确定需要存储的实体和实体的属性。实体是指需要存储的具体事物，属性是指实体的特征或描述信息，如某个二手房为实体，而该二手房的房价、面积和户型等特征为属性。在确定实体和属性之后，需要确定实体之间的关系。

常见的实体关系有一对一、一对多和多对多关系，下面给出了用户查询数据、系统展示信息、系统预测房价和用户注册发送验证码的 E-R 图，如图 3-7 所示：

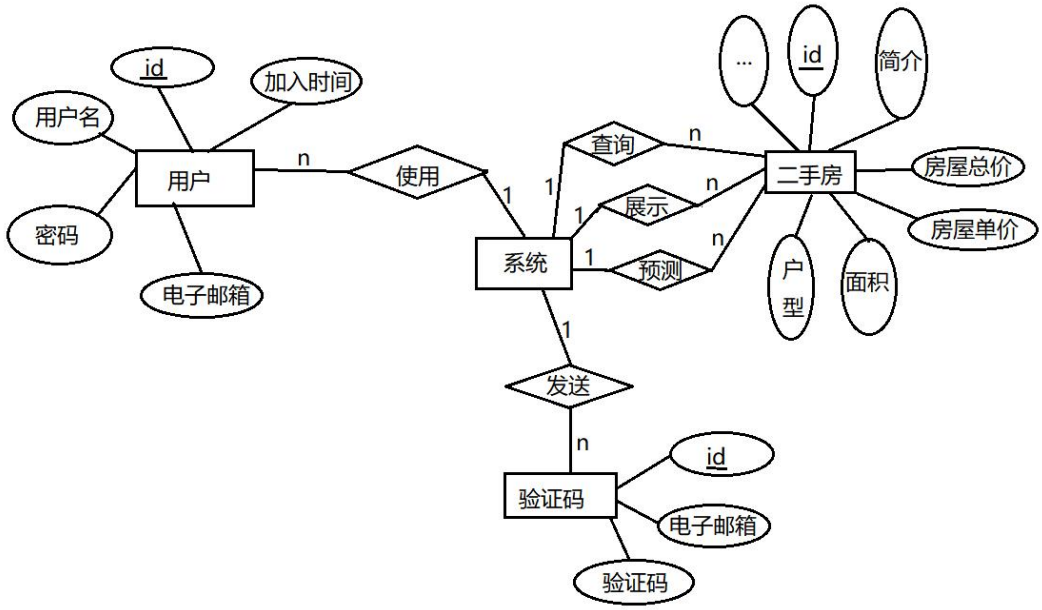


图 3-7　数据库实体关系图

表 3-1 为数据库中具体的表以及每个表的说明信息，其中 sh_data、forecast 表中所存储的是二手房信息，user、email_captcha、findpassword_captcha 表中所存储的是用户信息，alembic_version 表是利用 flask-migrate 库做数据库迁移默认生成的表。

表 3-1　数据字典

序号	表名称	说明
1	sh_data	经过简单的数据清洗后存入的二手房数据信息
2	forecast	对某些特征处理后的二手房数据信息
3	user	用户注册身份信息
4	email_captcha	注册-邮箱验证码
5	findpassword_captcha	找回密码-邮箱验证码
6	alembic_version	数据库迁移版本

（3）物理设计：根据需求选择合适的数据库软件，常用的数据库软件有 Oracle、MySQL、SQL Server、PostgreSQL 等，

Oracle 性能优越，安全性强，但属于商业软件，更适合大型企业级网站开发，复杂度和成本都较高；

MySQL 成本低、性能好、开源学习难度小，非常适合小中型网站开发，对海量数据的处理能力有限；

SQL Service 开发效率高、安全性高，但属于商业软件，成本较高，不太适合中小型网站应用；

PostgreSQL 安全性好，开源成本低，但性能较差，不太适合处理大量数据。

根据本系统数据的特点与各数据库的优缺点，最终选择 MySQL 数据库作为本系统的数据存储工具。在明确数据库软件之后，建立系统专属的数据库并取名为“secondhouse1”，根据逻辑分析所绘制的 E-R 图，创建数据库中的表，包括表名、字段、数据类型与长度，将数据获取与数据处理后得到的数据做好分类后存储在不同表中，数据表分别为 sh_data、forecast、user、email_captcha、findpassword_captcha，其中列别名、字段名与数据类型如下表 3-2，表 3-3，表 3-4，表 3-5，表 3-6 所示。

表 3-2　二手房信息 sh_data 表

列别名	字段名	数据类型
编号	id	int
标题	title	varchar(255)
房屋总价	total_price	float
房屋单价	unit_price	int
房屋面积	square	float
户型	size	varchar(255)
楼层	floor	varchar(255)
房屋朝向	direction	varchar(255)
房屋类型	type	varchar(255)
县市区	district	varchar(255)
所在区域	nearby	varchar(255)
小区	community	varchar(255)
装修情况	decoration	varchar(255)
电梯有无	elevator	varchar(255)
梯户比例	elevatorNum	varchar(255)
房屋性质	ownership	varchar(255)

表 3-3　二手房信息处理后 forecast 表

列别名	字段名	数据类型
编号	id	bigint
标题	title	text
房屋总价	total_price	double
房屋单价	unit_price	bigint
房屋面积	square	double
楼层	floor	text
房屋朝向	direction	text
房屋类型	type	text
县市区	district	text
所在区域	nearby	text
小区	community	text
装修情况	decoration	text
电梯有无	elevator	text
梯户比例	elevatorNum	text
房屋性质	ownership	text
户型处理后	dataSize_new	bigint

表 3-4　用户身份信息 user 表

列别名	字段名	数据类型
编号	id	int
用户名	username	varchar(100)
密码	password	varchar(200)
电子邮箱	email	varchar(100)
注册时间	jointime	datetime
手机号	phone	varchar(200)
通讯地址	address	varchar(200)
性别	sex	varchar(200)
年龄	age	varchar(200)

表 3-5　注册-邮箱验证码 email_captcha 表

列别名	字段名	数据类型
编号	id	int
电子邮箱	email	varchar(50)
验证码信息	captcha	varchar(50)

表 3-6　找回密码-邮箱验证码 findpassword_captcha 表

列别名	字段名	数据类型
编号	id	int
电子邮箱	email	varchar(50)
验证码信息	captcha	varchar(50)

4. 系统实现、测试与部署

4.1　数据获取

要开发一个二手房数据系统，最关键的第一步便是原始数据信息的获取。根据前期需求分析，确定了链家二手房交易网站（wf.lianjia.com）作为目标数据的来源网站，具体数据获取的实现过程如下：

（1）明确数据获取需要 python 爬虫技术，本系统的爬虫部分主要需要 requests 库和 lxml 库实现。

（2）搜索得到链家二手房交易网站的主域名，然后进入该网站查询潍坊市的二手房信息，得出该网站中各网址的跳转变化规律，即 URL 固定部分+变化数字。

（3）设置多个 USER_AGENTS，利用 randint 方法随机获取 UA 来模拟随机浏览器操作，避免目标网站封锁 IP。

（4）自定义一个爬虫类 SpiderFunc，类中定义一个 spider 函数用于存放爬虫程序。

（5）spider 函数首先通过 requests.get 方法对目标 URL 发送请求并得到响应，然后利用 etree.HTML 方法解析 HTML 文档，再通过 xpath 方法设置特定搜索范围来获取文档中的节点信息，这便完成了网站列表页的信息爬取。

（6）本系统需要得到二手房的各种详细信息，而列表页所展示的信息有限，因而还需要完成各二手房详情页的数据爬取。在（5）的基础上再使用 xpath 方法获取文档中跳转详情页的 URL，然后再利用 requests.get 方法对详情页 URL 发送请求并获得响应，利用 etree.HTML 方法解析 HTML 文档，再通过 xpath 方法获取二手房的主题、房价、单价、户型、面积等所需信息，并将这部分信息作为键值汇集到新创建的 item 字典中。

（7）在 Spider_wf.py 中定义 write_csv 函数和 write_db 函数，这两个函数的作用是将 item 字典内容分别写入到 CSV 文件和数据库中。write_csv 函数的实现方式是利用 open 方法打开本地新建的 CSV 文件，使用 csv.DictWriter 方法与 writerow 方法将 item 字典按照预先设定好的字段有序写入；write_db 函数的实现方式是先利用 pymysql.connect 方法连接数据库，再使用 get 方法获取字典中各个键所对应的值，然后利用 execute 方法执行 insert 数据库操作，最后使用 commit 方法提交事务，将数据真正传入数据库。

（8）由于链家二手房网站中潍坊市各县市区展示的二手房数量不等，有的县没有二手房信息，有的县仅有一千多条二手房信息，有的县最多三千条二手房信息，因而在实现程序中还需要利用 if-else 语句对该网站各县市区页数做判断。

（9）在爬虫程序长时间大量爬取数据时，可能会出现连接中断，使用 try-except 作异常处理可以解决。当出现异常时，程序随机睡眠 15 到 30 秒，睡眠结束后继续执行页面解析。

通过执行上述爬虫程序，最终获取到坊子区、潍城区、奎文区、寒亭区、青州市、寿光市、经济技术开发区和高新技术产业开发区八个潍坊县市区的共计 20826 条二手房数据信息，并分别存储到“二手房数据.csv”和数据库 sh_data 表中。

4.2 系统实现

开发二手房数据分析预测系统前首先要明确需要用哪种 Web 应用程序模型来完成系统，Web 应用程序模型常见的有两种：B/S 模型（Browser/Server）与 C/S 模型

（Client/Server）。

前者如图 4-1 所示，B/S 模型是基于 Web 浏览器和 Web 服务器的网络应用程序模型，应用程序的处理逻辑在服务器端，浏览器作为客户端向服务器发起请求，服务器响应请求并返回 HTML、CSS、JavaScript 等 Web 页面元素，浏览器将这些元素组装成可视化的 Web 页面并呈现给用户，这种模型适合需要广泛分发和访问的网站；

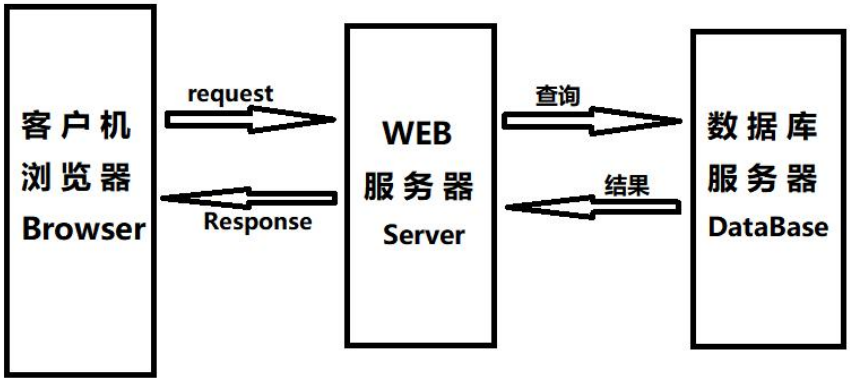


图 4-1　B/S 模型结构图

后者如图 4-2 所示，C/S 模型是基于客户端和服务器的网络应用程序模型，应用程序的处理逻辑在客户端和服务器端共同完成，客户端和服务器通过网络进行通信，完成数据传输和处理，这种模型更适合功能要求更高的多媒体应用、游戏等。

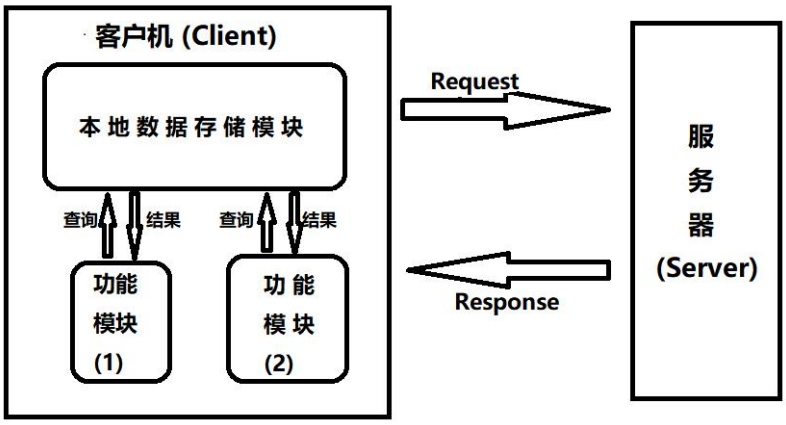


图 4-2　C/S 模型结构图

根据 C/S 模型的原理可以看出，C/S 模型的应用程序将部分数据缓存到本地客户端软件，对网络带宽要求较低，可以减轻对服务器大数据量访问的负载，但缺点也十分明确，它需要单独开发、部署和更新客户端软件，实现与测试需要的成本较高，开

发难度大，同时对于不同的操作系统，还要开发不同版本的软件，不易实现跨平台，通用性较差；相较而言，B/S 模型的应用程序依赖于用户使用浏览器对服务器完成请求工作，这种模型虽然对服务器和网络带宽要求较高，功能较为有限，但较低的开发部署成本使得应用程序完成周期短，测试与部署难度小，另外强大的跨平台性从用户角度出发，用户不需要花费多余的精力来下载软件，节省了时间成本与物力成本，这更有利于系统前期的推广与普及。结合本系统的特点与两种应用程序模型的适用性，系统选择 B/S 模型开发二手房数据分析网站，由用户在浏览器上完成对二手房数据分析预测系统的操作。具体功能模块与页面设计过程如下：

4.2.1 登陆注册

（1）注册页

用户身份信息记录模块指的是收集用户个人信息并与登录身份做比对，用户在首次使用本系统时可以前往注册界面注册个人账号，需要注册的信息有电子邮箱、邮箱验证码、用户名、密码、确认密码，如图 4-3 所示。

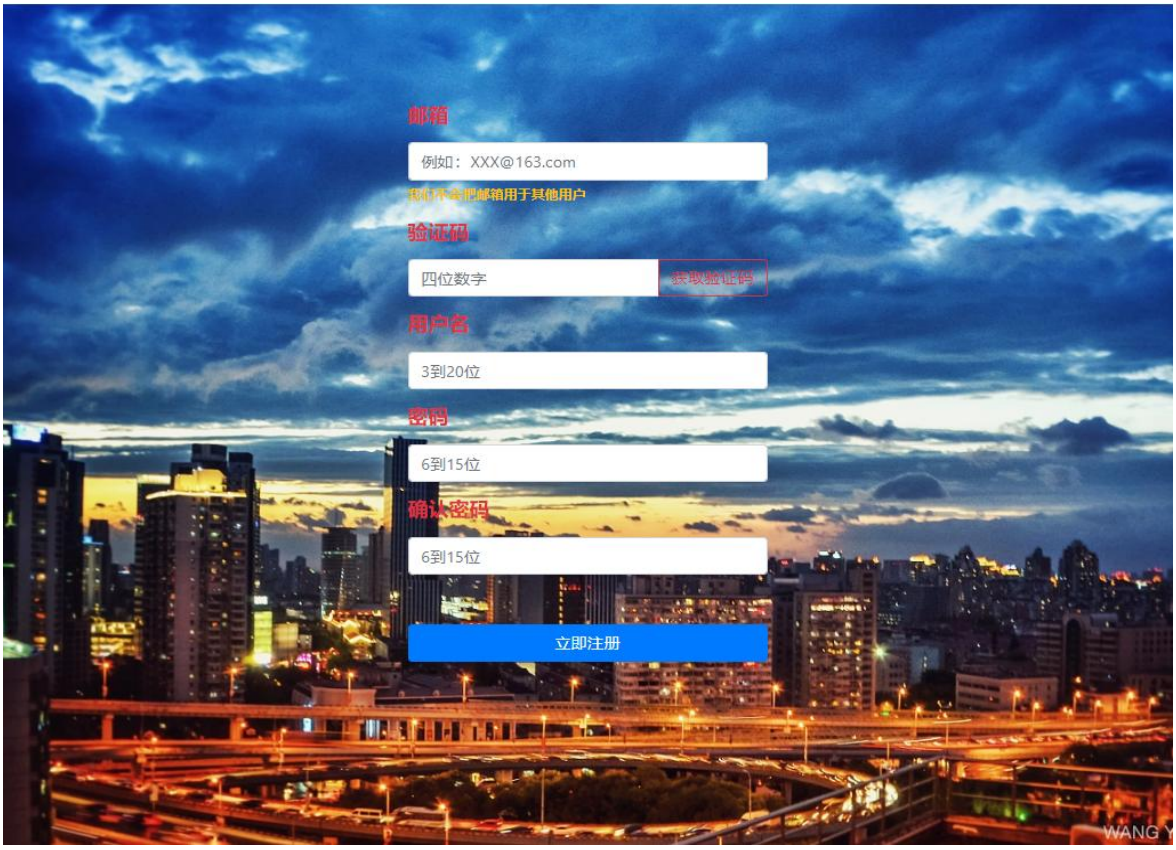


图 4-3　注册页面图

在注册时首先需要填写电子邮箱，然后点击“获取验证码”按钮，系统会在 Flask 框架后端执行 user.py 蓝图文件其中的视图函数 get_email_captcha，随后该函数会利用 config 配置文件中的系统邮箱配置信息，通过 mail 模块向注册者填写的邮箱发送网站注册验证码信息，该验证码中的四位数字部分是通过 random.randint 模块生成的随机数，确保每次发送的验证码都不重复，同时，验证码信息会写入到实现创建的数据库 email_captcha 表用于后续的验证工作，效果图如 4-4 所示。

图 4-4　注册验证码发送成功图

在按照图 4-3 中输入框提示信息注册完成后如果输入内容不符合系统要求，会提示输入内容有误，重新填写后点击立即注册，系统会把符合要求的用户信息利用 generate_password_hash 模块对用户密码进行 md5 加密后存储到数据库中，同时收集了用户名、邮箱、加入时间等信息，如图 4-5 所示。

id	use	password	email	jointime
1	roo	pbkdf2:sha256:260000	2547985435@	2023-02-16 23:10:37
2	littl	pbkdf2:sha256:260000	1716311167@	2023-02-17 13:40:00
3	bey	pbkdf2:sha256:260000	1941450030@	2023-03-05 13:40:56

图 4-5　用户身份信息测试图

（2）登录页

在成功注册后，系统会自动跳转到登录页面，登录页如图 4-6 所示。

图 4-6　登录页面图

按照注册收集的信息，该页面登录需要填写注册完成的电子邮箱和密码，其中利用 check_password_hash 模块将输入的密码内容与数据库中存取密码的哈希值检查，当用户输入完成后，若内容有误，系统会提示输入内容有误且当前页不会发生变化；若信息无误，当前页则跳转到该系统网站的首页并且系统由非登陆状态转为登陆状态，如图 4-7 所示。

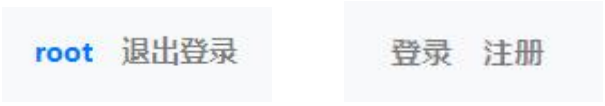


图 4-7　登陆状态与非登陆状态图

（3）用户资料页

在登录状态下，点击网站左上角蓝色的用户名可以跳转到用户个人资料界面，如图 4-8 所示，该页面由用户基本信息与用户信息修改两部分组成，其中用户基本信息中的用户名、性别、年龄、邮箱、手机号、地址和注册日期等数据都是通过数据接口从后端传递过来的，在用户信息修改部分中利用 form 表单提交数据并由后端视图函数 message 接收后对数据库 user 表中的相应字段的值做修改。

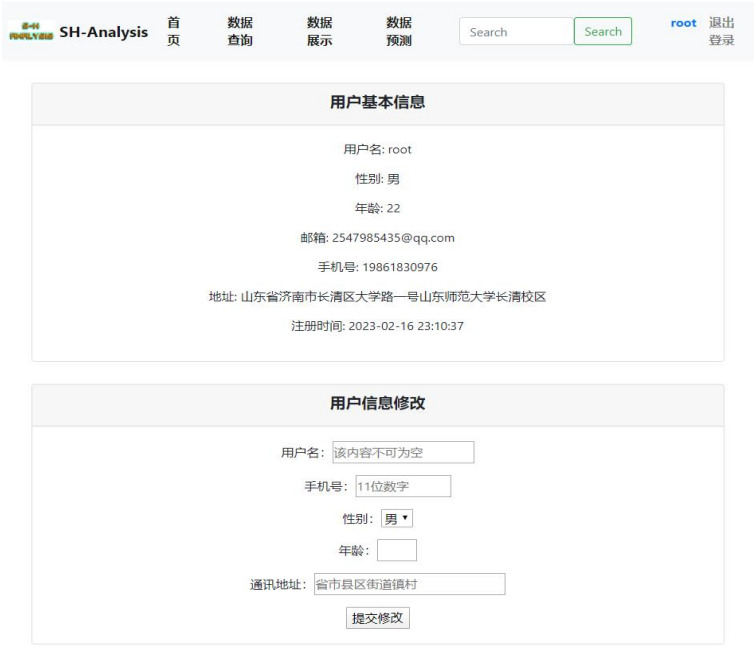


图 4-8　个人资料页面图

（4）找回密码页

当用户忘记密码时可以点击图 4-6 登录页中的“忘记密码？”来找回密码，界面

效果如图 4-9 所示。

图 4-9　找回密码页面图

该页面中用户需要按要求输入邮箱、验证码、新密码与确认密码后才能完成修改密码。其中在输入已注册的邮箱后，点击获取验证码，具体实现过程同（1）中获取验证码类似，获取的验证码存储到数据库 findpassword_captcha 表中，效果图如 4-10 所示。

图 4-10　找回密码验证码发送成功图

在修改完密码并提交后，如果内容有误，系统会在当前页提示某项输入错误；如果内容无误，则会提示“密码修改成功”。

4.2.2 首页界面实现

由于在系统界面设计中提前绘制了网站首页的页面布局，因而在实际开发中，只需要确定用哪种具体样式即可。

如图 4-11 所示，首页采用了单元模块化设计方式，首先利用 Bootstrap 框架中的 Hero 大块屏组件对系统网站的名称与功能模块做介绍，其次利用 Bootstrap 框架的栅格组件将网页布局划分为左右两部分，左侧给出了“二手房示例”单元（部分二手房的简介、图片与详情页跳转）、“二手房讯息”单元（某些关于二手房的最新媒体消息）与“词云展示”单元（利用 python 词云程序+二手房数据生成特定二手房信息词云）；右侧给出了“政策公告”单元（房地产及二手房行业的相关政策公告）。

图 4-11　首页效果图

4.2.3 数据检索查询

当用户需要查看二手房信息并需要搜索特定地区、小区的二手房数据时，可以使用系统网站的数据查询模块。如图 4-12 所示，数据查询页面中包括大标题、搜索部分和数据显示部分。其中数据显示部分利用了前端框架 layUI 中的数据表格样式与前端框架 Bootstrap 的分页组件。数据表格中的数据来源于数据库 sh_data 表，由于二手房数据信息较多，无法显示在一个页面中，因而需要对数据表格做分页处理，通过对页面显示效果与用户体验度的考量和对后端路由节点的理解，本页面设计每五十条数据信息制成一个数据表格，不同数据表格放在不同网址下，不同网址的子节点为 page 索引而不需要单独创建新的 HTML 文件。在 JavaScript 设置分页逻辑，不同网址之间的跳转功能通过分页组件完成。搜索部分由信息输入框与提交按钮组成，信息输入框共有三个，功能分别为关键词查询、地区查询与小区查询，当在输入框输入内容后，点击提交按钮，表单数据会以“POST”方法提交给后端接受“POST”方法的路由，该

路由下的视图函数利用 request 请求读取特定输入框的文本内容，并将文本内容与数据库中数据进行模糊匹配，找到含有搜索内容的数据信息，视图函数将该数据返回到前端数据表格并显示出来，这样数据查询功能便实现了。

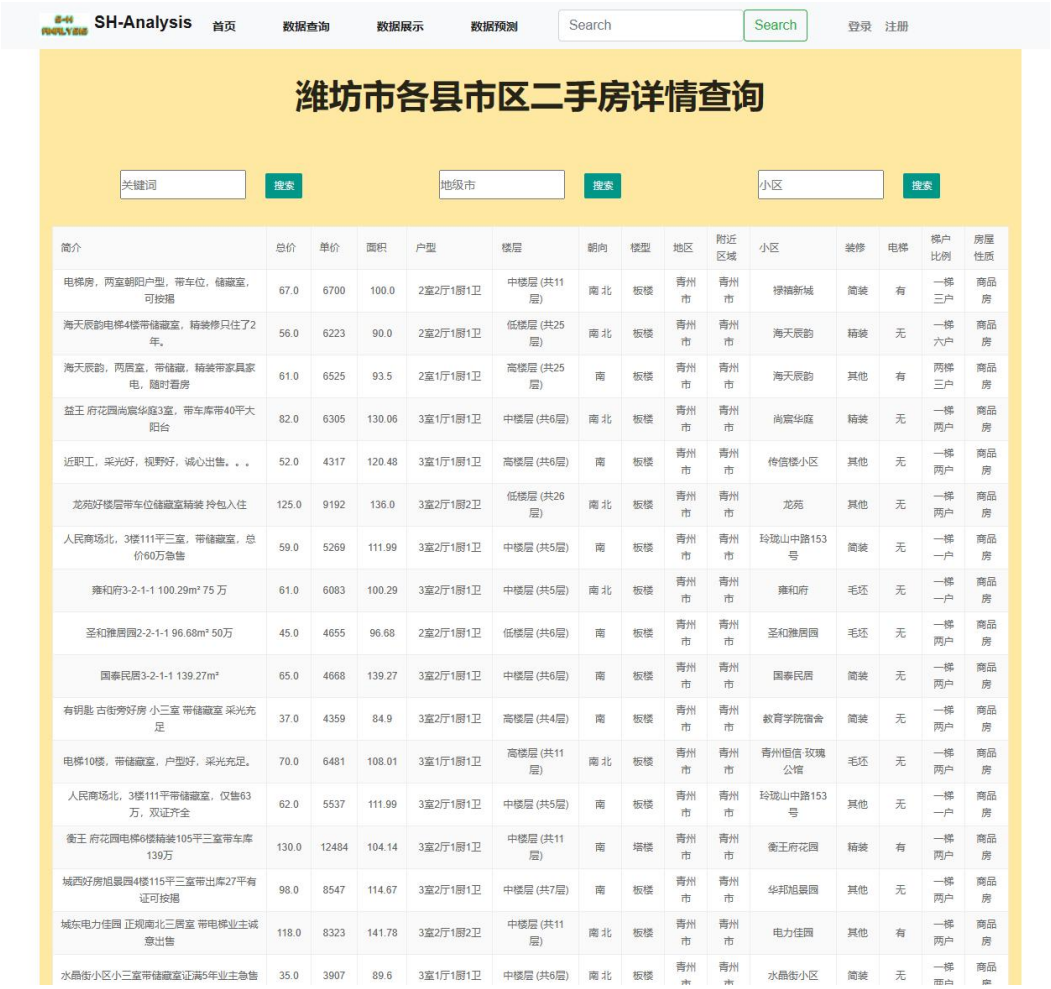


图 4-12　信息检索页面效果图

4.2.4 数据可视化

数据可视化是数据分析中一项最基本的实现形式，它通常的表现形式为将纷繁错杂的文本型数据信息转化为清晰明确、有规律的图像型数据信息，它可以帮助人们从复杂的信息中找到各种数据特征间的关系，建立起复杂数据间的联系，从而发现某些数据的规律性。该系统便根据前期需求分析的结果从数据可视化功能出发，搭建各种不同特征间的数据图表。

如图 4-13 所示为数据可视化功能模块的导航页，同样利用单元模块化设计页面，整体布局分为了三大版块：数据可视化简介，图表展示和地图预览。

图 4-13　数据可视化导航页效果图

首先数据可视化简介单元对数据可视化这一名词做了简要介绍，其次图表展示单元利用了 Bootstrap 框架中的卡片组，对不同的可视化图表分为三类，分别用卡片对内容做简要介绍并提供跳转到图表详情页的链接，所有图表有柱状图类、饼图类、折线图类、散点图类，这些图表的绘制利用了 echarts 组件库，整个过程是在前端 JavaScript 中确定具体图形信息；在后端中利用 numpy、pandas 等数据分析库对二手房数据做预处理、特征提取、数据转换和分组整合，输出数据为前端数据接口所需要的形式并将其传递给前端。最终所绘制的可视化图表如下：

（1）柱状图类：该类图表有基础柱状图如图 4-14、折柱混合图如图 4-15、大数据量柱图如图 4-16 和 3D 柱图如图 4-17，其中图表介绍部分利用了 Bootstrap 组件中的 POP 提示，点击按钮会显示图表详情内容，再点击一次详情信息会隐藏。

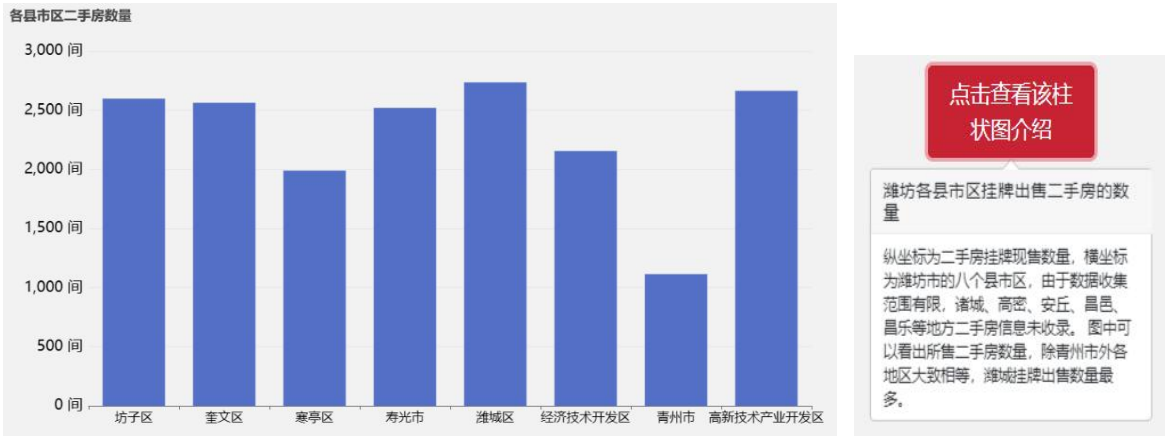


图 4-14　基础柱状图及图表介绍

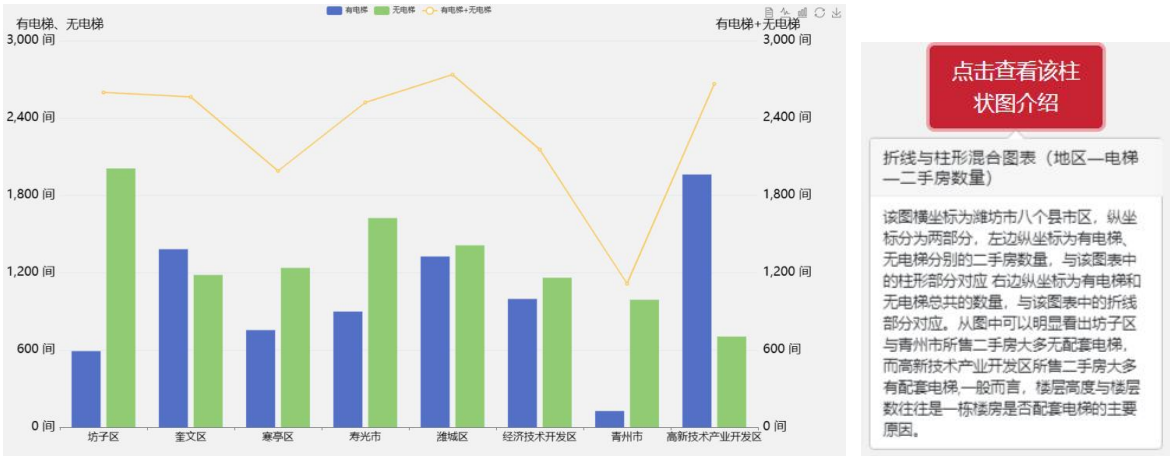


图 4-15　折柱混合图及图表介绍

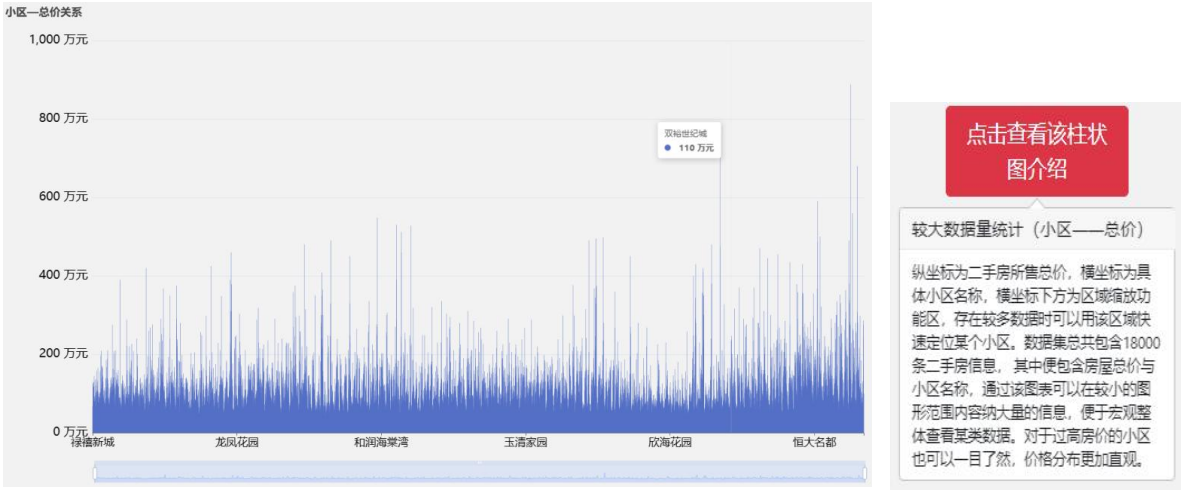


图 4-16　大数据量柱图及图表介绍

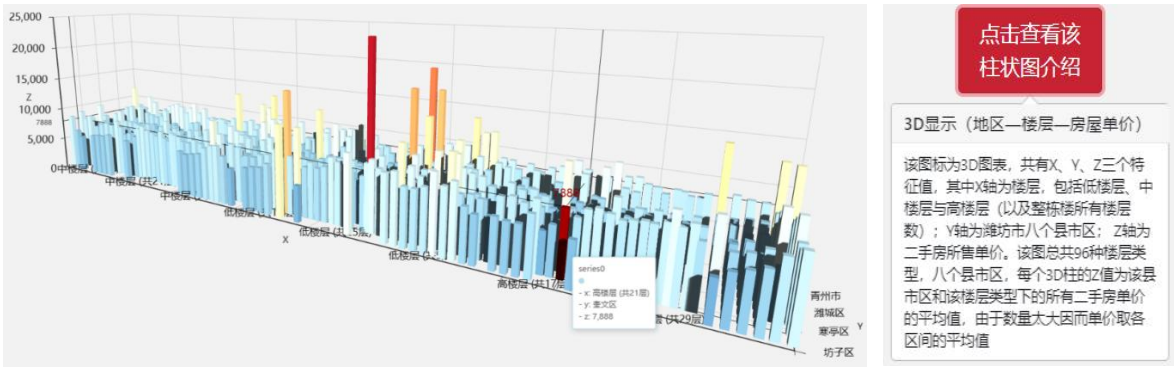


图 4-17　3D 柱状图及图表介绍

（2）饼图类：该类图表有基础饼图如图 4-18、朝向—纹理饼图如图 4-19、楼型—纹理饼图如图 4-20、户型—纹理饼图如图 4-21 和装修情况—纹理饼图如图 4-22。

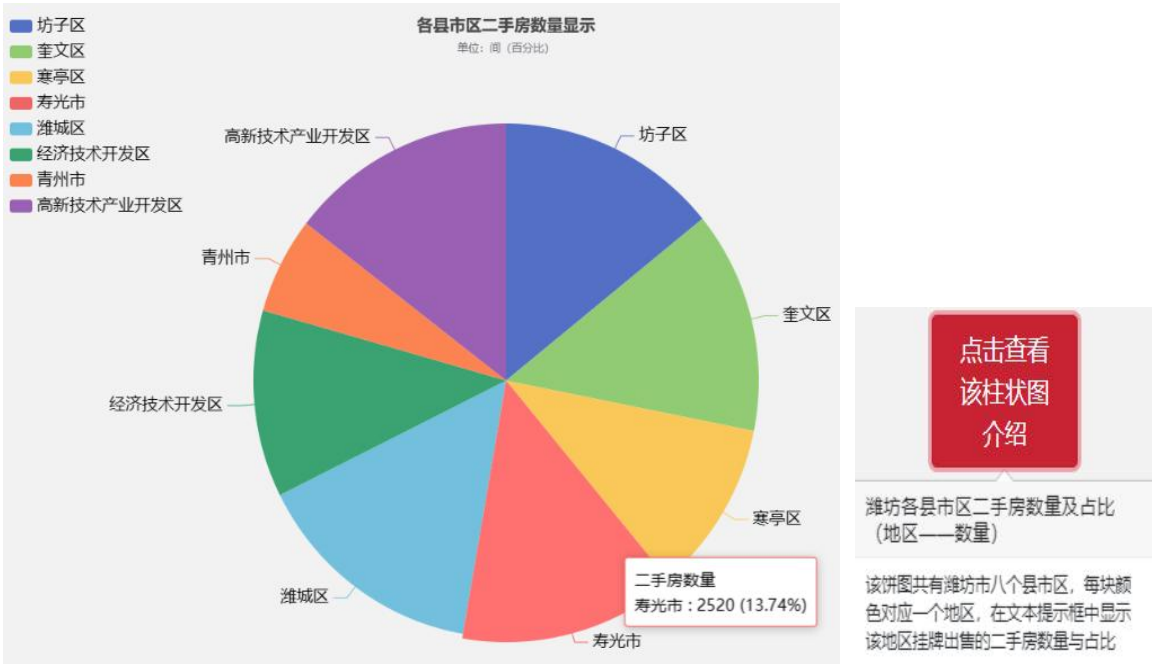


图 4-18　基础饼图及图表介绍

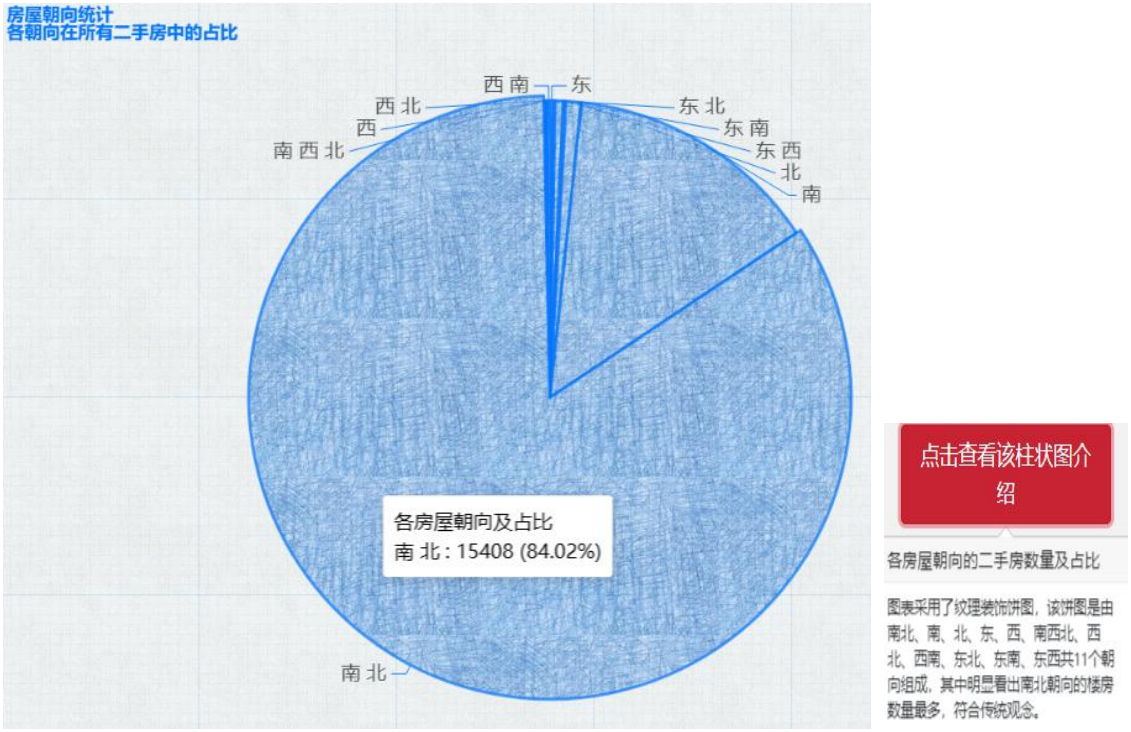


图 4-19　朝向—纹理饼图及图表介绍

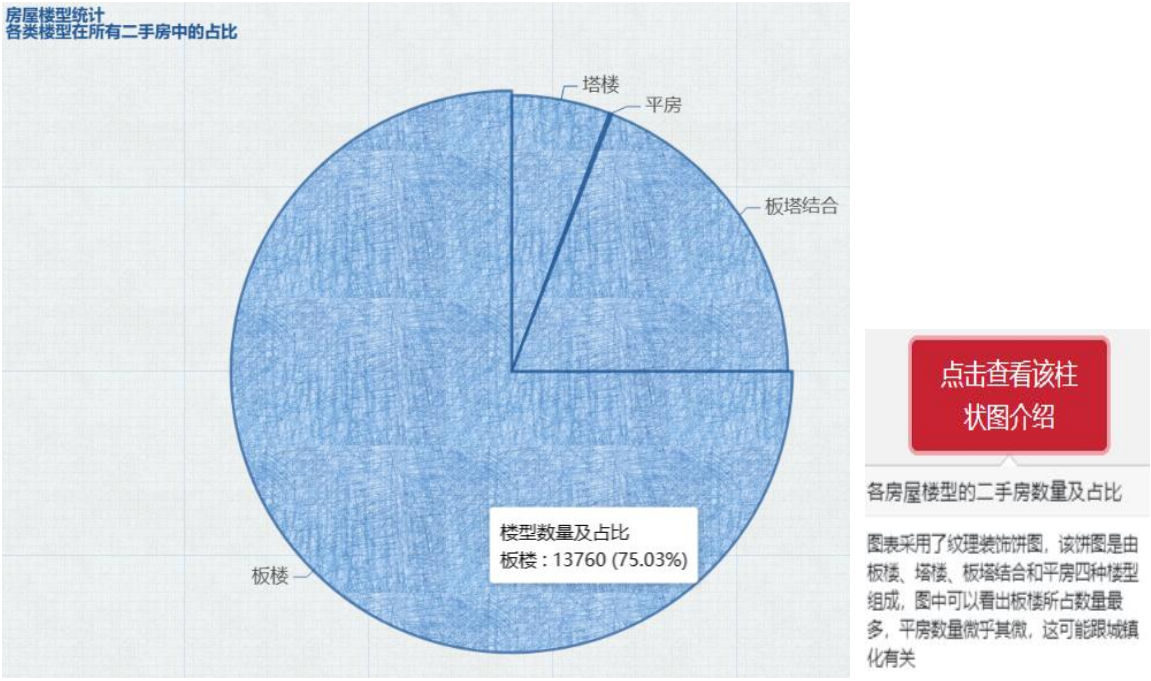


图 4-20　楼型—纹理饼图及图表介绍

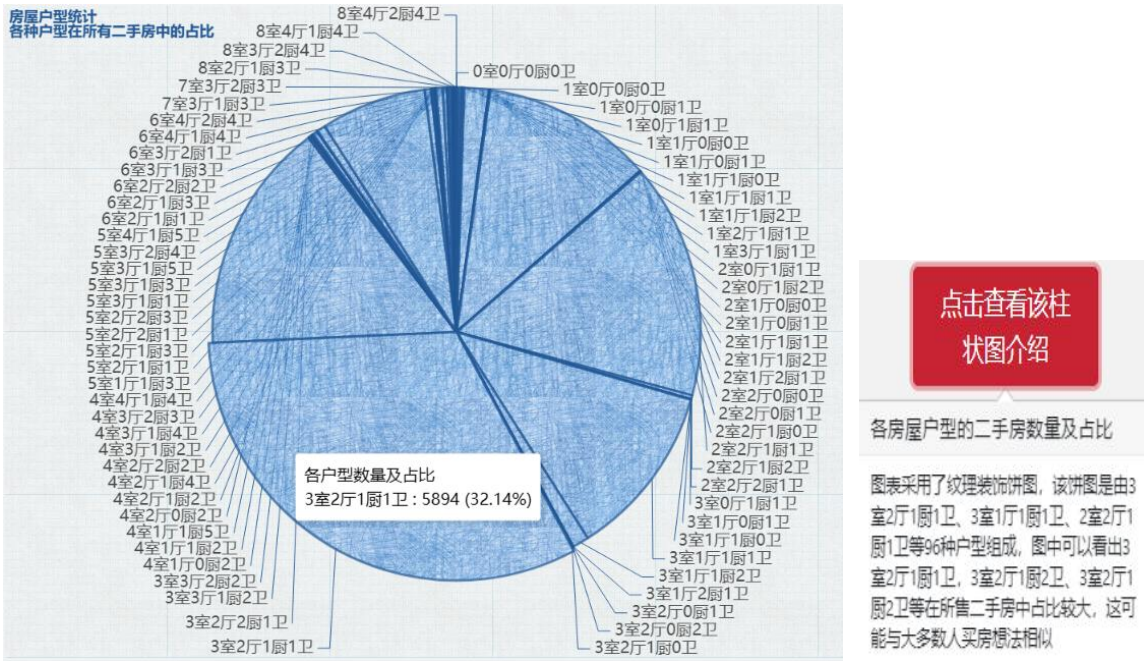


图 4-21　户型—纹理饼图及图表介绍

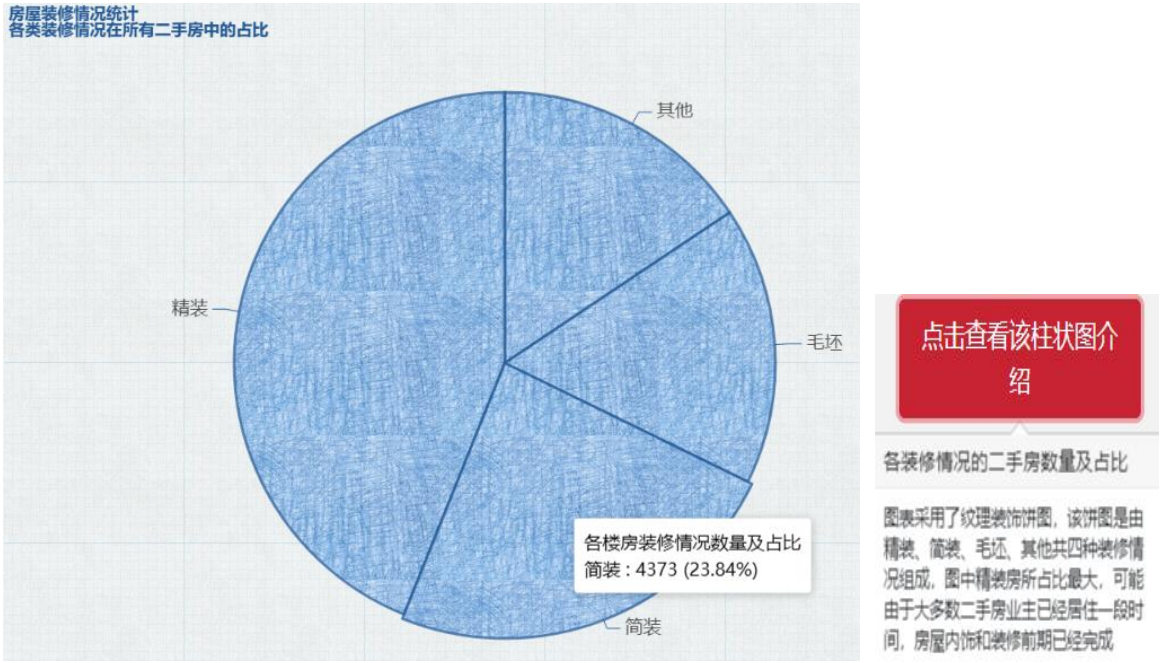


图 4-22　装修情况—纹理饼图及图表介绍

（3）折线图类：这类图表有叠加折线图如图 4-23，该图为多条折线叠加而成，点击图例可以查看指定折线。

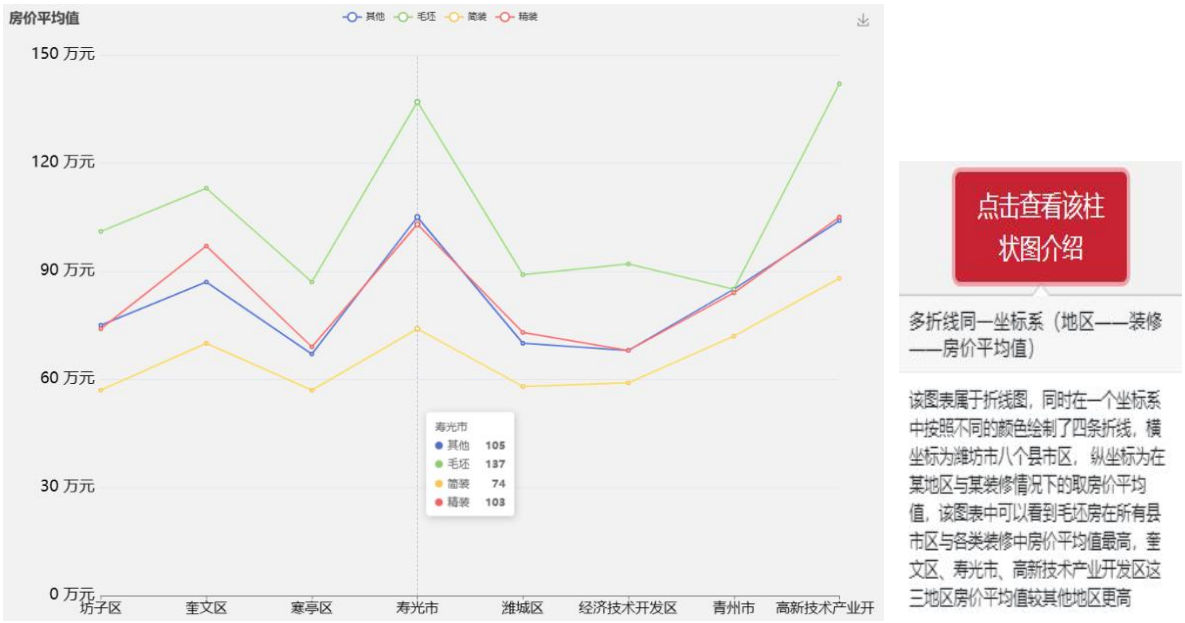


图 4-23　叠加折线图及图表介绍

（4）散点图类：这类图表有流式渲染和视觉映射式散点图如图 4-24。

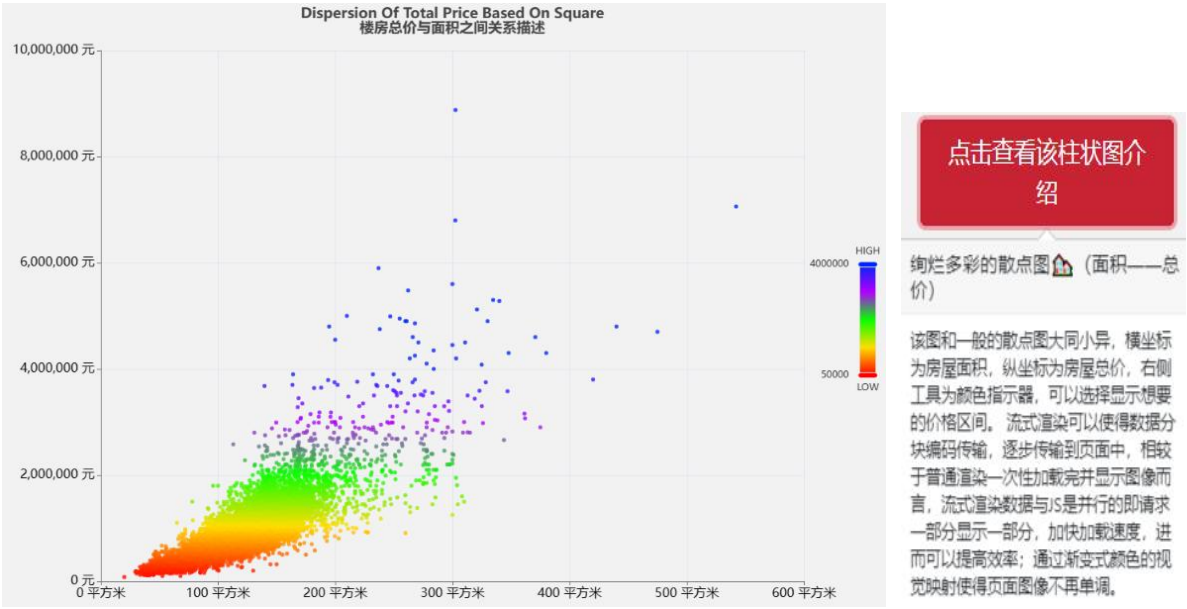


图 4-24　流式渲染和视觉映射式散点图及图表介绍

数据可视化模块最后一个单元为地图预览，如图 4-25 所示，它可以帮助用户对二手房所在地区、小区等位置信息进行查询，该版块包括地点搜索框、地址显示框与地图模块。地图模块通过调用百度地图 API 将卫星地图组件引入到网页中，并在 JavaScript 中设置地图控件实现对地图的管理；地点搜索功能是指在搜索框输入具体地点之后，利用百度地图 API 中的 BMap 模块搜索位置功能把当前搜索点在地图中展示出来；此

外，随机在地图点击一处，JavaScript 会捕捉到“click”条件，执行 getLocation 方法获取点击位置的地址，并将其显示到地址显示框。

图 4-25　地图预览效果图

4.2.5 数据预测

数据分析工作除了对已有数据做特征处理与归纳总结外，还需要能给出已知信息来推出未知信息，数据预测便是利用数据挖掘中的某些分类回归算法建立模型，根据已有数据信息得出可能的数值。

如图 4-26 所示，数据预测模块导航页包含方法简介与房价预测两个版块，前者描述了数据预测需要做哪些任务以及每项任务的目的与方法；后者利用 Bootstrap 框架中的水平排列卡片对所采用的预测模型做了简要介绍并给出跳转链接。

图 4-26　数据预测导航页图

本系统分别利用线性回归与 KNN 两种算法模型对潍坊市二手房数据做预测，两种模型具体实现效果图如下：

（1）线性回归模型：该模型具体为简单线性回归模型，分别利用一元线性回归与多元线性回归从面积、房间数量、装修情况、楼型、地区、电梯、房屋性质等要素对房价做出预测。

如图 4-27 所示为一元线性回归模型所在页面的效果图，左侧图片是利用 Matplotlib

可视化库对模型中的训练数据做可视化处理，右侧包含信息输入部分与预测结果输出部分，输入框的限定输入内容为二手房面积，单位为平方米，数据类型为 int 类型。

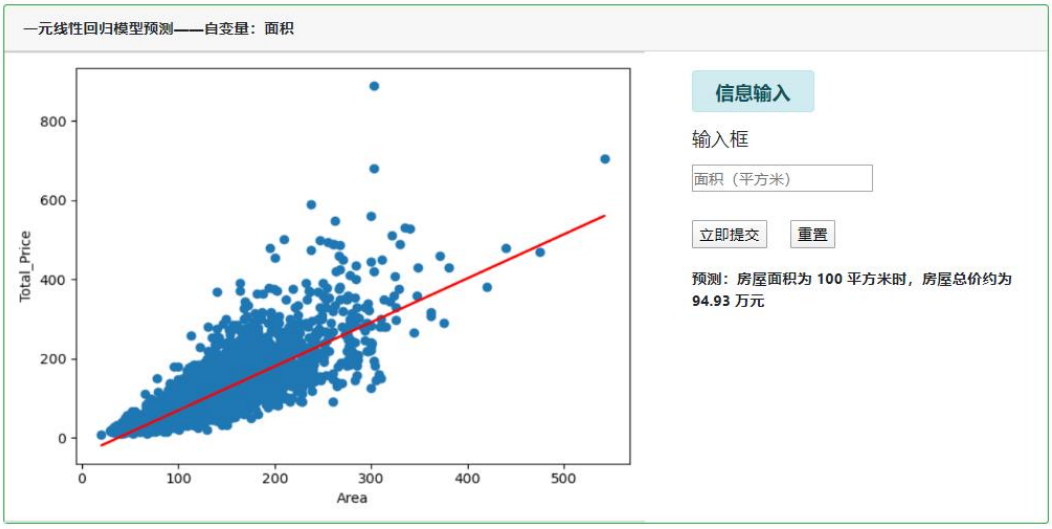


图 4-27　一元线性回归模型预测页面图

当在输入框输入内容完成，点击“立即提交”按钮后，系统前端 form 表单会以 POST 方法将输入框文本内容传递给后端接收 POST 方法路由下的视图函数，视图函数会对表单的提交按钮做条件判断，如果按钮名称为“立即提交”，便将文本内容交给写好的一元线性回归函数 Simple_linear_regression 做预测。

Simple_linear_regression 函数中利用 pandas 库对数据库 forecast 表做特征提取并将房价数据与房屋面积数据分组成不同的 DataFrame，然后利用了 Scikit-learn 机器学习库中的 LinearRegression 方法来建立模型并用面积—房价数据训练模型，训练完成的模型可以利用 predict 方法对外界输入的数据做房价预测，最后返回预测结果并通过 {{ simple }}传值给前端页面。

如图 4-28 所示为多元线性回归模型所在页面的效果图，左侧图片同样是模型可视化结果，右侧同样为上方的信息输入部分与下方的预测结果展示部分组成，与一元线性回归不同的是，多元线性回归所需要的数据不仅仅是面积，而是面积、房间数量、装修情况、楼型、地区、电梯、房屋性质共七个元素，其中面积、房间数量都是输入框形式，输入类型都为 int 类型，房间数量限定为 1-20，其余元素则采用下拉菜单形式，装修情况有其他、毛坯、简装和精装，楼型有塔楼、板楼、板塔结合和平房，地区有坊子区、潍城区、寒亭区、奎文区、青州市、寿光市、经济技术开发区和高新技术产

业开发区，电梯类型分有和无，房屋性质有商品房和房改房。

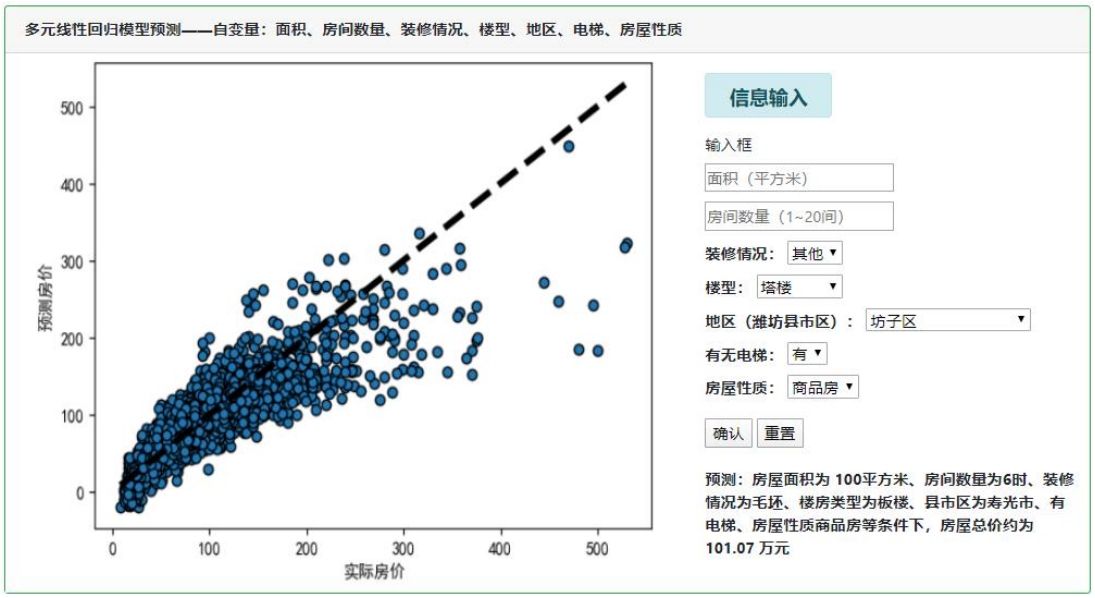


图 4-28　多元线性回归模型预测页面图

具体实现过程为当在输入框按限定条件输入内容并在下拉菜单选择合适条件后，点击“确认”按钮，前端 form 表单传递过程和一元线性回归模型预测类似，主要区别为根据表单提交按钮名称为“确认”，后端会将文本信息加工交给多元线性回归函数 Multiple_linear_regression 做预测。

Multiple_linear_regression 函数与 Simple_linear_regression 函数不同之处在于输入参数数量与类型不同，Multiple_linear_regression 函数的参数共七个，其中分 square_guess、dataSize_new_guess 为数值型数据，而 decoration_guess、type_guess、district_guess、elevator_guess 和 ownership_guess 为离散型数据，所以获得数据之后要对 forecast 表中数据与输入数据两者中的离散型数据做重编码处理，这里选择的是 LabelEncoder 标签编码方法，它可以将离散型文本数据编码成连续的整数。

在编码完成后从 forecast 表选择出特征为'square'、'dataSize_new'、'decoration'、'type'、'district'、'elevator'和'ownership'的自变量 DataFrame 与特征为'total_price'的因变量 DataFrame，将两组数据划分出训练集与测试集，并同样利用 LinearRegression 方法建立模型并用训练集训练模型，把测试集输入模型以获得测试性预测数据并采用 r2_score 方法计算模型评估指标——R2 决定系数，所得 R2 分数为 0.72284，这表示模型拟合准确度较高，最后利用 predict 方法得到对应输入内容的预测结果，使用{{ multiple }}传

值给前端。

简单线性回归是数据预测最常使用的一种算法模型，方法简单、拟合度较好，但是由于模型容易受到极端值的影响，对源数据要求比较高，预测结果可能出现异常或偏差值较大，因而本系统又采用了 KNN 模型做补充对照。

（2）K 近邻模型：如图 4-29 所示，页面布局同样左侧为模型可视化图片，右侧为信息输入部分与预测结果部分，信息输入部分又新增了楼层要求与梯户比例两个输入框，输入类型为文字，具体输入内容可以参考右侧示例。

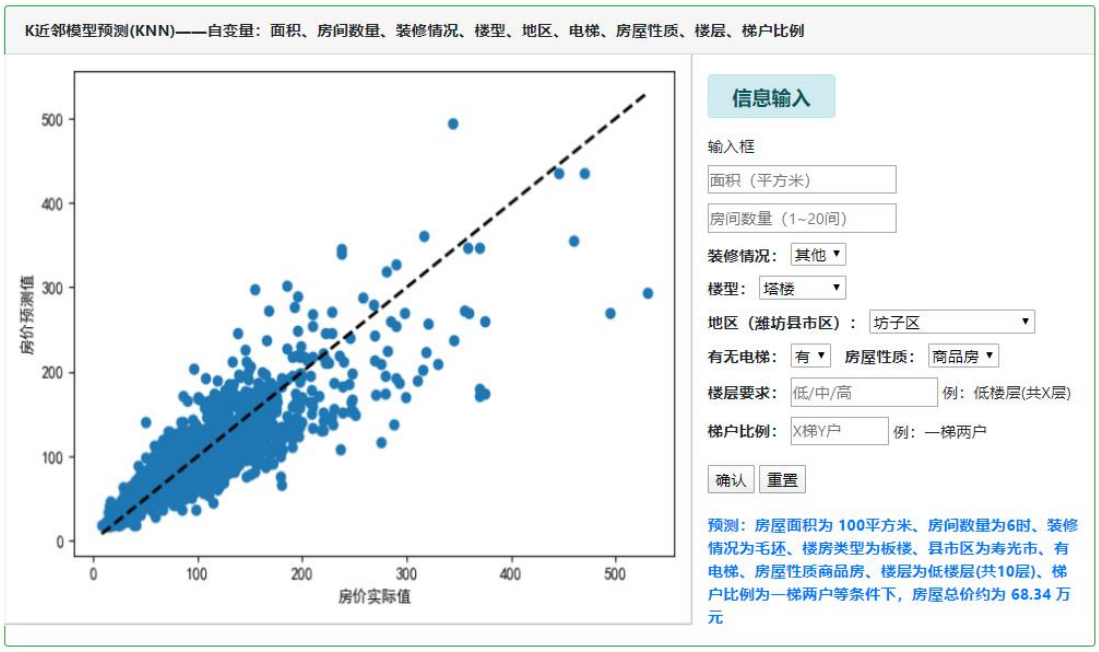


图 4-29　K 近邻模型预测页面图

预测实现过程为当用户输入完内容点击“确认”按钮后，前端 form 表单会以 POST 方法将输入信息传递给后端接收 POST 方法的 knn 视图函数，视图函数中利用 get 方法获取到的数据传入函数 KNN。KNN 函数中有面积、房间数量、装修情况、楼型、地区、有无电梯、房屋性质、楼层和梯户比例共九个参数，同样利用 pandas 库对 forecast 表中的源数据分组，拆分为特征变量（'square'、 'dataSize_new'、'decoration'、'type'、'district'、'elevator'、'ownership'、'floor'、'elevatorNum'）与目标变量（'total_price'），分别按照一定比例划分训练集与测试集，利用 Scikit-learn 库中的 KNeighborsRegressor 方法建立 K 近邻模型并用训练集训练模型，根据测试结果评估模型的 RMSE、MAE 以及 R2 指标，计算可得 RMSE=23.15496，MAE=14.58415，R2_score=0.74819，最后

利用 predict 方法得到对应输入内容的预测结果，使用{{ knn }}传值给前端。

K 近邻算法通过利用不同的 K 值对待测数据做分类，得出待测数据所属类别及该类别目标变量的预测值，这种方法对于极端值的处理效果比较好，通过简单线性回归与 K 近邻两个模型的评估指标对比和二者预测结果与现实情况之间的差异度，可以得出 K 近邻模型在处理大量数据时拟合度与准确率要优于简单线性回归模型。

4.3 系统测试与部署

在系统开发完成之后，可能会有某些方面不能满足用户需要，还需要开发人员或测试人员从不同的角度来测试系统各种性能与功能。本节内容主要包括两部分，一是对系统功能性与非功能性的测试，二是在系统网站完成之后如何实现其他用户访问该二手房数据分析网站。

4.3.1 测试环境

表 4-1 介绍了本机的开发及测试环境，为测试不同浏览器对于操作本系统网站是否有影响，本测试选用了五种主流浏览器。

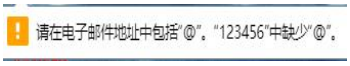
表 4-1　系统测试工具

测试工具	工具信息
个人电脑	I5-8300H，16G 内存，256G 系统盘，1T 数据盘，OS：Windows 10
校园网	100M 无线网，网络协议：IPv4
浏览器	Chrome，Microsoft Edge，Firefox，星愿浏览器，QQ 浏览器，360 浏览器

4.3.2 测试结果

如表 4-2 所示，测试主要内容为用户登陆注册测试，数据查询测试，数据预测测试，数据可视化测试和非功能模块测试。用户登录注册模块中注册测试项有注册用户邮箱、用户验证码、用户名和密码，登录测试项有登录用户邮箱与密码；数据查询模块中测试内容为搜索功能与分页显示；数据预测中对不同条件测试是否能得出预测结果；数据可视化模块中测试项为地图功能；非功能模块主要针对各网页间的跳转情况做测试。

表 4-2　测试行为及对应结果

测试序号	测试项	测试行为	测试结果
1	注册页测试	邮箱：输入 123456	
2		邮箱：XXX@qq.com（已注册）	注册页提示“该邮箱已被注册”
3		验证码：输入 1234	注册页提示“邮箱或验证码有误”
4		用户名：不到 3 位或超出 20 位	注册页提示“用户名格式错误”
5		密码：不到 6 位或超出 20 位	注册页提示“密码格式错误”
6	登录页测试	邮箱：YYY@qq.com（未注册/不存在）	登录页提示“邮箱不存在或未注册”
7		密码：123456（不存在）	登录页提示“密码错误”
8		邮箱、密码：不填写	登录页提示“密码错误”
9	数据查询测试	搜索：关键词输入框搜索“寿光市”	数据表格显示标题中含“寿光市”的二手房
10		分页：点击数据表格下方分页页码	“上一页”“下一页”及具体页码按钮均可以使用，并能完成网址的遍历
11	数据预测测试	一元线性回归模型：输入框输入 120	**预测：房屋面积为 120 平方米时，房屋总价约为 117.17 万元**
12		多元线性回归模型：面积 200，七个房间，其他，塔楼，坊子区，有电梯，商品房	**预测：房屋面积为 200平方米、房间数量为7时、装修情况为其他、楼房类型为塔楼、县市区为坊子区、有电梯、房屋性质商品房等条件下，房屋总价约为 196.82 万元**
13		KNN 模型：面积 150，七个房间，精装，板楼，奎文区，有电梯，商品房，中楼层(共 15 层)，一梯两户	**预测：房屋面积为 150平方米、房间数量为7时、装修情况为精装、楼房类型为板楼、县市区为奎文区、有电梯、房屋性质商品房、楼层为中楼层(共15层)、梯户比例为一梯两户等条件下，房屋总价约为 126.56 万元**
14	数据可视化测试	地图预览：输入测试地点“山东师范大学长清校区”	

15		地图预览：随机点击潍坊市范围内地点	显示位置：山东省, 潍坊市, 坊子区, 凤凰大街, 7号 显示位置：山东省, 潍坊市, 奎文区, 玉清东街,
16	非功能性测试	点击各页面中所含的跳转链接	各类链接均可以跳转到目标页面

4.3.3 服务器部署

要让更多的用户访问本系统网站，就需要将项目部署到互联网中，这一过程有两种实现形式，一种形式是可以选择个人电脑作为服务器，配置防火墙和路由器端口转发，若网站项目使用 web 服务器，如 Nginx 或 Apache，还需要单独配置该服务器程序端口在本地计算机的端口，这种方法需要本地计算机长时间开启并且存在信息泄露的风险，安全性差，另一种形式是使用云服务器。

本文采用云服务器部署的方法，选用阿里 ECS 服务器，服务器规格为 2vCPU、2G 内存、40G 系统盘、1Mbps 带宽和操作系统 CentOS 7.9，搭载宝塔 Linux 面板，需要单独安装 Nginx 服务器软件、PHP 7.4、phpMyAdmin5.2 与 MySQL5.7 等。

将系统项目上传到 ECS 中，项目路径为/www/wwwroot/secondhouse，利用宝塔面板中的 python 项目管理器插件，安装项目所需的 python3.9 环境，并且在项目管理中添加 python 项目，如图 4-30 所示。

图 4-30　服务器添加项目图

点击确认按钮后，python 项目管理器会根据项目文件夹中的 requirements.txt 内容去安装项目所需要的依赖包（requirements.txt 需要提前利用终端命令 pip3 freeze > requirements.txt 来生成），项目添加完成后，配置 gunicorn 文件中的本地服务器域名端口，修改为 127.0.0.1:5656，然后需要项目映射到准备好的域名，并将服务器 IP 与端口添加到站点域名中，如图 4-31 所示。

图 4-31　绑定服务器 IP 地址

由于系统网站的数据查询、可视化与预测功能模块都需要二手房数据信息，而在上述配置过程中数据库中数据为上传，所以需要单独建立数据库，设置数据库名称、用户名和密码，并将本地电脑中数据库的数据导入到云服务器中，具体实现方式为通过服务器提供的 phpMyAdmin 工具（基于 web 的 MySQL 数据库管理工具）创建网站需要的数据表，然后将本地 SQL 数据文件导入到 phpMyAdmin 中，数据库便创建完毕。

整个过程操作完成之后，项目便部署完成，通过在浏览器输入 8.130.26.250:5000 便可以访问本系统网站。

在对网站实际访问时发现，网页加载速度较慢，用户等待页面显示所需时间较长，通过使用浏览器的开发者工具查看 Network 选项发现，服务器调用静态资源 echarts.js 与 echarts-gl.js 文件花费大量时间，如图 4-32 所示。

图 4-32　优化前网站载入时间

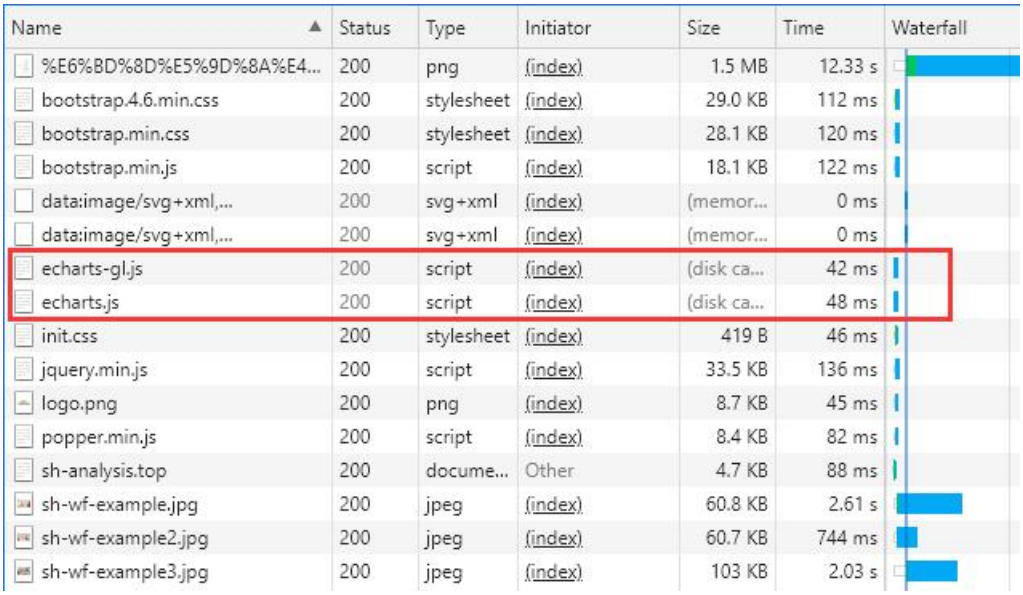
对于解决网站加载 JS 静态资源用时过长而导致网页空白这一问题，主要有以下三种解决方法：

1、利用 CDN 内容分发网络使用户根据自己网络的特点更快、更稳定地访问到节点服务器中的资源，从而加快网页加载速度。

2、将服务器 base.html 文件 head 标签中的 script 标签放到 body 标签中，实现在页面完全显示之后再加载 JS 文件，但实质上并未解决 JS 资源加载时间长的问题。

3、给 script 标签设置延迟脚本或异步脚本，实现浏览器先下载 JS 文件，再延迟执行。

由于互联网提供了 echarts 及其 3D 库的 CDN 加速服务，因而在服务器端 HTML 文件中使用 bootCDN 来加载 echarts 与 echarts-gl 资源，最后通过测试发现，网页加载时间明显加快，如图 4-33 所示。



Name	Status	Type	Initiator	Size	Time	Waterfall
%E6%BD%8D%E5%9D%8A%E4...	200	png	(index)	1.5 MB	12.33 s	
bootstrap.4.6.min.css	200	stylesheet	(index)	29.0 KB	112 ms	
bootstrap.min.css	200	stylesheet	(index)	28.1 KB	120 ms	
bootstrap.min.js	200	script	(index)	18.1 KB	122 ms	
data:image/svg+xml,...	200	svg+xml	(index)	(memor...	0 ms	
data:image/svg+xml,...	200	svg+xml	(index)	(memor...	0 ms	
echarts-gl.js	200	script	(index)	(disk ca...	42 ms	
echarts.js	200	script	(index)	(disk ca...	48 ms	
init.css	200	stylesheet	(index)	419 B	46 ms	
jquery.min.js	200	script	(index)	33.5 KB	136 ms	
logo.png	200	png	(index)	8.7 KB	45 ms	
popper.min.js	200	script	(index)	8.4 KB	82 ms	
sh-analysis.top	200	docume...	Other	4.7 KB	88 ms	
sh-wf-example.jpg	200	jpeg	(index)	60.8 KB	2.61 s	
sh-wf-example2.jpg	200	jpeg	(index)	60.7 KB	744 ms	
sh-wf-example3.jpg	200	jpeg	(index)	103 KB	2.03 s	

图 4-33　优化后网站载入时间

5. 总结

Python 作为一门简单、高效的语言，它提供了许多功能强大的数据分析库与交互式工具，像 numpy、pandas、Matplotlib、sklearn 和 Jupyter Notebook 等，正是这样使得 python 在数据分析领域扮演重要角色，而房地产或二手房交易向来与数据分析联系紧密，因而本文结合对二手房交易现状与潍坊市二手房信息的调查，建立了一个基于 python 的二手房数据分析预测系统，该系统能够帮助用户获取潍坊二手房详情信息，快速分析二手房市场的地区差异、价格走势以及不同房屋条件下的房价预测。本文所做的主要工作如下：

（1）对课题研究背景和意义做了介绍，明确了本文的研究对象、研究目标和现实

意义，参考国内外有关二手房数据分析的资料与文件，汇总整理了二手房数据分析预测系统的国内外研究现状。

（2）在确定研究方向以后，需要对系统所需要具备的功能或性能开展实际调查，通过对收集到的二手房购房者需求作分析，形成了本文的数据查询、可视化展示、预测等功能性需求与更新性、安全性、界面样式等非功能性需求。

（3）系统设计环节在总体上描述了系统所需要完成的任务及每项任务的流程顺序，从系统功能、系统界面与数据库三个方面分别阐述，确定了通过 python+Flask 框架搭建网站，利用绘图软件提前绘制好页面样式草图，搭建了 MySQL 数据库存储环境。

（4）讲述了原始数据来源及获取数据的具体实现过程，对实际数据爬取工作中遇到的问题作了阐释并给出解决方法。

（5）对系统搭建实现过程作了详细介绍，其中关键实现有数据查询模块的数据表格与分页、数据可视化的 echarts 图表、数据预测的简单线性回归模型与 K 近邻模型，给出了一元线性回归、多元线性回归与 K 近邻做房价预测的实际运行原理。

（6）测试了搭建的系统网站的各项功能并将系统网站部署到服务器，实现公共网络访问本网站。

虽然本系统在整体上达到了符合需求分析的预期效果，但有些部分不太完整或存在不足，还需要继续完善，主要表现如下：

（1）数据查询模块中的地区搜索不太准确，实际输入某个地区时会搜索字段为标题的一项而不是字段为地区的一项。

（2）数据预测模块中的算法模型在预测房屋面积过小条件下的房价时，所得到的预测值可能为负值或者 0，但是如果将预测值与偏差值直接相加，可能使最终预测结果偏大，准确率降低，还需要进一步研究算法模型的改进或者考虑建立其他算法模型作对比。

参考文献

[1] David Gray. House Price Diffusion: An Application of Spectral Analysis to the Prices of Irish Second-Hand Dwellings[J]. Housing Studies, 2013, PP 869-890.

[2] Jon Stobart. Domestic textiles and country house sales in Georgian England[J]. Business History, 2019, PP 17-37.

[3] Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez, V. Raul Perez-Sanchez. Determinants of the Price of Housing in the Province of Alicante (Spain): Analysis Using Quantile Regression[J]. Sustainability, 2019, PP 437.

[4] Christian A.L. Hilber, Olivier Schöni. On the economic impacts of constraining second home investments[J]. Journal of Urban Economics, 2020,Volume 118.

[5] Koktashev Vladislav, Makeev Vladimir, Peresunko Pavel. Comparison of prices depending on factors in the secondary housing market[J]. SHS Web of Conferences, 2021, Volume 116.

[6] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000(04): 54-57.

[7] 王薇. 信息时代的视觉信息图表设计[J]. 装饰, 2006, (04): 128-129.

[8] 董倩, 孙娜娜, 李伟. 基于网络搜索数据的房地产价格预测[J]. 统计研究, 2014, 31(10): 81-88.

[9] 原继东, 王志海. 时间序列的表示与分类算法综述[J]. 计算机科学, 2015, 42(03):1-7.

[10]冷建飞, 高旭, 朱嘉平. 多元线性回归统计预测模型的应用[J]. 统计与决策, 2016, No.451(07): 82-85.

[11]孔钦, 叶长青, 孙赟. 大数据下数据预处理方法研究[J]. 计算机技术与发展, 2018, 28(05):1-4.

[12]王自成, 朱家明, 陈华友. 基于逐步回归筛选的回归组合预测模型[J]. 统计与决策, 2019, 35(17): 75-78.

[13]徐志, 金伟. Python 爬虫技术的网页数据抓取与分析[J]. 数字技术与应用, 2020, 38(10).

[14]钟机灵. 基于 Python 网络爬虫技术的数据采集系统研究[J]. 信息通信, 2020, No.208(04).

[15]崔明明, 刘晓亭, 李秀婷, 董纪昌. 数据特征驱动的房地产市场集成预测研究[J]. 管理评论, 2020, 32(07): 89-101.

[16]李函谕, 魏嘉银, 卢友军. 基于随机森林的深圳二手房价格预测与分析[J]. 现代信息科技, 2021,

5(15): 100-104.

[17]姬正骁. 基于 Python 的武汉二手房信息爬取及分析[J]. 信息与电脑（理论版）, 2022, 34(16): 195-199.

[18]洪丽华, 黄琼慧. 基于 Python 爬虫技术的研究[J].价值工程, 2022, 41(34): 154-156.

[19]任妮, 吴琼, 栗荟荃. 数据可视化技术的分析与研究[J].电子技术与软件工程, 2022, No.234(16):180-183.

[20]韦依洋, 吴一凡, 李永远. Python 技术在数据可视化中的应用研究[J].福建电脑, 2022, 38(01): 27-31.

[21]曾悠. 大数据时代背景下的数据可视化概念研究[D].浙江大学, 2014.

[22]宋健. 基于集成学习的二手房产数据分类研究[D].西南交通大学, 2018.

[23]徐阳阳. 济南市二手房市场分析及价格预测[D].山东师范大学, 2019.

[24]唐铭昊. 山东省潍坊市二手房价格分析[D].山东师范大学, 2022.

[25]仲姝锜. 基于机器学习的数据预处理框架研究[D].西安工业大学, 2022.

[26]朱俊. 二手房数据分析系统的设计与实现[D].西南交通大学, 2018.

[27]武空军. 济南市二手房交易管理系统的设计与实现[D].电子科技大学, 2013.

致谢

当写完这篇论文的时候，大学生活很快就要结束了，在这四年的大学时光里，学到了很多，收获了很多。

感谢我的论文导师魏艺老师能让我选择她的课题并且非常负责地指导我论文写作，给我指出修改意见，解答写论文遇到的疑惑。

感谢家人对我生活学习的关心与激励，给予我精神与物质的支持，鼓励我不断进步。

感谢辅导员苗老师对我日常生活的关怀和帮助。

感谢我的舍友和朋友们在学习与生活上给予我的帮助，为我的大学生活增添了不少乐趣。

感谢评阅老师与答辩组老师对我论文的指导与建议。

在我即将毕业的时候，我由衷的感谢山东师范大学信息科学与工程学院给了我一个美好的大学生活。

山东师范大学本科生毕业论文（设计）题目审批表

学院（部）：信息科学与工程学院 (章)　　时间：2022 年 11 月 15 日

<table>
<tr><td rowspan="5">课题情况</td><td>题目名称</td><td colspan="5">基于 Python 的二手房数据分析预测系统</td></tr>
<tr><td>课题类型</td><td colspan="5">设计-应用研究</td></tr>
<tr><td>教师姓名</td><td>魏艺</td><td>职称</td><td>讲师</td><td>学位</td><td>博士</td></tr>
<tr><td>课题来源</td><td colspan="5">其他</td></tr>
<tr><td>指导教师审批意见</td><td colspan="6">同意

指导教师签名：魏艺
2022 年 11 月 15 日</td></tr>
<tr><td>系或教研室审批意见</td><td colspan="6">同意

负责人签名: 鲁燃
2022 年 11 月 15 日</td></tr>
<tr><td>学院(部)审批意见</td><td colspan="6">同意

学院（部）（盖章）
2022 年 11 月 15 日</td></tr>
</table>

山东师范大学
本科生毕业论文（设计）开题报告

题　　目：基于 Python 的二手房数据分析预测系统

学院(部)：　信息科学与工程学院

专　　业：　计算机科学与技术（公费师范生）

姓　　名：　常子儒

学　　号：　201911990102

指导教师：　魏艺

2022 年 12 月 25 日

一、选题的性质

设计-应用研究

二、选题的目的和意义

目的：设计基于 Python 的二手房数据分析预测系统，通过 Python 爬虫抓取各地二手房相关数据，数据分析功能实现对二手房多年数据变化的分析，算法分析未来变化的预测。

意义：随着社会经济迅速发展，中国城镇化建设加速，二手房市场迅猛发展，交易量居高不下。二手房既是住房，也可以被用以理财投资，因此人们对于二手房房产价格及相关信息评估的需求也随之增大。

该选题旨在设计二手房信息分析预测模型，并通过该数据分析预测系统，实现二手房房产信息数据的预处理，利用数据可视化处理帮助人们更清晰的认识各种复杂信息，进而从数据中发现有用信息，接着通过机器学习、数据挖掘算法对二手房数据进行分类预测，为二手房信息变化提供合理的参考模型，这不仅可以对消费者进行价格引导，使二手房购房者和投资者了解二手房的价格区间,对购房与投资等有一定的引导意义，也可以给政府制定二手房相关政策提供一定的数据支撑，有实用意义。

三、与本课题相关的国内外研究现状，预计可能有所创新的方面

1、国外研究现状

国外在很早的时候就有专门从事二手房交易的职业--二手房中介与房屋经纪人，相关企业起步早，软件技术发达，有完善的二手房信息管理软件。

1993 年初 Matthew Gray’s Wandered 在麻省理工学院开发出第一个网络爬虫，开启了数据采集自动化的进程，为二手房数据收集提供了便利。2013 年 David Gray 运用谱分析的方法确定了房价的动态变化，研究是否可以在爱尔兰城市的二手住宅房地产市场上找到房价连锁效应的证据[1],2019 年 Jon Stobart 借鉴北安普敦郡 1761-1836 年拍卖记录，以确定二手纺织品数量和性质变化，以及涨价与估价方式，并揭示了国家房屋拍卖是二手商品再循环中的一个关键机构[2],2019 年 Raul-Tomas Mora-Garcia 等人分析和量化了阿利坎特市场上二手房的要价与表征它们的属性之间的关系，结果表明，阿利坎特市

场的价格分割，北部沿海地区的价格高于南部和内陆[3]，2020 年 Christian A.L. Hilber，Olivier Schöni 调查了自然舒适度高的地方对于二手房投资者政治限制的影响，利用实验“瑞士第二家园计划”得到研究结果，季节性旅游地点对二手房市场的发展做出限制[4]。2021 年 Koktashev Vladislav 等学者对二手房市场价格取决因素进行了比较，利用机器学习的非参数方法,构建预测模型，层次聚类等方法，实现了公寓成本预测的高精度,揭示和描述了二手住房对象价格形成的特殊性[5]。

2、国内研究现状

国内二手房数据分析开始研究时间与国外相比相对较晚。

2000 年刘明吉，王秀峰和黄亚楼针对当年随数据库与人工智能发展起来的新兴学科——数据挖掘做出来解释，将对数据处理的研究放于数据挖掘研究工作的重点[6]。随后数据挖掘中数据预处理研究也逐渐增加：2014 年以董倩为代表的学者以北京、上海、天津、重庆等 16 个大中城市的二手房价格和新房价格为研究对象，以来自我国最大搜索引擎的百度搜索指数为数据基础，实现了数据处理与分析，采用了支持向量机（SVM）与随机森林对采集到的数据做了回归预测，并进一步拟合[8],在数据挖掘领域，2015 年原继东与王志海介绍了关于时间序列表示方面的非数据适应性表示方法、数据适应性表示方法和基于模型的表示方法，针对时间序列的分类方法,着重介绍了基于时域相似性、形状相似性和变化相似性的分类算法,并对未来的研究方向进行了进一步的展望[9]。随着计算机技术发展，大数据时代下数据组织模式与类型多样化，数据质量参差不齐等使得数据感知表达、理解计算面临巨大挑战，2018 年以孔钦为代表的学者分析了数据预处理的主要任务，总结了对“脏数据”的几种常用处理方法，阐释了数据在清洗、集成、变换与归约过程中的常用算法[11]。在数据采集方面，爬虫技术成为了一种强大的数据抓取工具，2020 年徐志，金伟通过阐释网络爬虫的原理，利用 Python 爬虫技术对网页数据进行抓取[13]，同年钟机灵开发基于 Python 网络爬虫技术的数据采集系统实现了主题数据的自动采集，利用 urllib、Beautiful Soup、threading 库设计开发了包含数据爬取、异常处理、robots 协议管理及多线程管理等模块的系统模型框

架[14]；2022 年姬正骁利用 Python 爬虫工具对链家网武汉市各行政区在售二手房数据进行采集，并对爬取到的信息进行清洗，最后使用 Matplotlib 和 Pyecharts 库进行可视化分析 [17]，同年洪丽华和黄琼慧从 Python、爬虫技术与网页爬虫三方面进行阐述爬虫如何帮助用户搜索整理相关数据[18]。在数据采集和预处理之后需要将数据更直观地呈现出来，2006 年王薇讲述了视觉信息图表将成为未来数据可视化的主流，通过对信息时代的视觉需求分析、视觉信息图表概念的阐述表达了信息时代视觉图表设计的存在价值[7]；2022 年任妮，吴琼，栗荟荃与韦依洋，吴一凡，李永远都探讨了数据可视化技术在 Python 工具的应用[19][20]。数据预测是数据挖掘领域更深层次的研究，多元线性回归算法既是数据挖掘中有效的一类算法又是机器学习中基础性的回归算法，2016 年冷建飞，高旭和朱嘉平将多元统计分析作为基础与前提，验证了相关结果改变对于多元线性回归方程整体的影响，并通过实例对模型做检验，提高准确度与效率，使原本回归结果得到最大程度上的优化[10]。组合预测相较于传统的单项预测而言，它能够有效集成各项预测方法的信息，在预测实践中有广泛的应用，2019 年王自成，朱家明和陈华友就组合预测模型方法改进这一问题，提出了逐步回归筛选的回归组合预测模型，改变自变量进入方程的方式，有效地将对实验结果不显著的变量筛选掉，并利用人口数据进行实例分析证明了筛选后的回归组合预测方法有更高的预测精度[12]。2020 年以崔明明为代表的学者采用组合集成预测的方法对房地产市场变化方向与水平进行了预测[15]，2021 年李函谕，魏嘉银和卢友军针对深圳市二手房市场房价预测问题，结合相关的八个特征变量，利用随机森林模型训练了房价预测模型，得出了二手房市场信息变动的结论[16]。

3、数据分析算法

（1）线性回归算法

线性回归算法用来解决回归问题，也就是预测连续值的问题，符合这种要求的数学模型被称为“回归模型”，按照自变量与因变量的数量关系分为一元线性回归分析和多元线性回归分析，根据变量之间是否为一次函数关系分为线性回归与非线性回归。

（2）逻辑回归算法

逻辑回归之所以比线性回归更加适合分类问题，因为逻辑回归在线性回归的基础上，将输出值 wT+b 通过 sigmoid 激活函数映射到[0,1]的区间，使得所得结果可以分为两类。

（3）集成多决策树的随机森林

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定，将一个输入样本进行分类，就需要将它输入到每棵树中进行分类。将若干个弱分类器的分类结果进行投票选择，从而组成一个强分类器，这就是随机森林算法的思想

4、参考文献

[1] David Gray.House Price Diffusion: An Application of Spectral Analysis to the Prices of Irish Second-Hand Dwellings[J].Housing Studies,2013,PP 869-890

[2]Jon Stobart.Domestic textiles and country house sales in Georgian England[J].Business History,2019,PP 17-37

[3]Raul-Tomas Mora-Garcia, Maria-Francisca Cespedes-Lopez, V. Raul Perez-Sanchez, Pablo Marti, Juan-Carlos Perez-Sanchez. Determinants of the Price of Housing in the Province of Alicante (Spain): Analysis Using Quantile Regression[J].Sustainability,2019,PP 437

[4] Christian A.L. Hilber,Olivier SchÖni.On the economic impacts of constraining second home investments[J].Journal of Urban Economics, 2020,Volume 118

[5]Koktashev Vladislav,Makeev Vladimir, Peresunko Pavel, Mikhalev Anton, Tynchenko Vadim.Comparison of prices depending on factors in the secondary housing market[J].SHS Web of Conferences,2021,Volume 116

[6]刘明吉;王秀峰;黄亚楼.数据挖掘中的数据预处理[J].计算机科学,2000,(04)

[7]王薇.信息时代的视觉信息图表设计[J].装饰,2006,(04)

[8]董倩;孙娜娜;李伟.基于网络搜索数据的房地产价格预测[J].统计研

究,2014,31(10)

[9]原继东;王志海.时间序列的表示与分类算法综述[J].计算机科学,2015,42(03)

[10]冷建飞;高旭;朱嘉平.多元线性回归统计预测模型的应用[J].统计与决策,2016,(07)

[11]孔钦;叶长青;孙赟.大数据下数据预处理方法研究[J].计算机技术与发展,2018,28(05)

[12]王自成;朱家明;陈华友.基于逐步回归筛选的回归组合预测模型[J].统计与决策,2019,35(17)

[13]徐志;金伟.Python 爬虫技术的网页数据抓取与分析[J].数字技术与应用,2020,38(10)

[14]钟机灵.基于 Python 网络爬虫技术的数据采集系统研究[J].信息通信,2020,(04)

[15]崔明明;刘晓亭;李秀婷;董纪昌.数据特征驱动的房地产市场集成预测研究[J].管理评论,2020,32(07)

[16]李函谕;魏嘉银;卢友军.基于随机森林的深圳二手房价格预测与分析[J].现代信息技术,2021,5(15)

[17]姬正骁.基于 Python 的武汉二手房信息爬取及分析[J].信息与电脑(理论版),2022,34(16)

[18]洪丽华;黄琼慧.基于 Python 爬虫技术的研究[J].价值工程,2022,(34)

[19]任妮;吴琼;栗荟茎.数据可视化技术的分析与研究[J].电子技术与软件工程,2022,(16)

[20]韦依洋;吴一凡;李永远.Python 技术在数据可视化中的应用研究[J].福建电脑,2022,38(01)

[21]曾悠.大数据时代背景下的数据可视化概念研究[D].杭州:浙江大学,2014

[22]宋健.基于集成学习的二手房产数据分类研究[D].成都:西南交通大学,2018

[23]徐阳阳.济南市二手房市场分析及价格预测[D].济南:山东师范大学,2019

[24]唐铭昊.山东省潍坊市二手房价格分析[D].济南:山东师范大学,2022

[25]仲姝锜.基于机器学习的数据预处理框架研究[D].西安:西安工业大学,2022

四、课题研究的可行性分析

1、课题可行性

当下房地产行业受政策影响产生波动较大，人们对于房地产领域一直在期房预售制与公摊面积两个问题上存在争议，期房在一定程度上缓解了开发商资金压力，刺激了房地产行业经济循环，为开发注入了新活力，但暴露的问题同样致命，与现房不同的是期房相当于购房者开盲盒，这种模式将风险从开发商转嫁到个人和银行，购房者承担了无法按期交房的风险，因而越来越多的购房者将目光转移到二手房交易上，这不需要承担房屋无法交付的风险，但是二手房的具体情况可能不是公开透明的，因而基于 Python 对二手房数据分析预测可以实现信息透明化，提供参考意见，具有实用意义，是可行的。

2、数据可行性

通过网络搜索已经收集了各地的二手房，小区均价，二手房单价，房屋户型，所在楼层，建筑面积，房屋公摊面积，房屋朝向，建筑结构，装修程度，电梯配套比例，这些数据对于研究二手房信息有作用。

3、方法可行性

Python 的爬虫功能可以短时间采集多个地区的二手房数据信息；通过对采集到的信息进行数据清洗，去除无关数据；进一步进行数据可视化处理，可以直观展示信息分析结果，令抽象、孤立的数据变得相互联系；最后是预测性数据分析，选择合适的机器学习算法。

五、课题研究的策略、方法和步骤

1、数据采集

（1）获取网页链接

选择合适的房产网站、二手房交易网站，观察需要爬取的网页变化规律，把获取的多个网页链接存入到字典中充当一个临时数据库，在需要用时直接通

过函数调用即可获得。

（2）数据存储

爬虫爬取到的网页将数据存入原始页面数据库，页面数据与用户浏览器得到的 HTML 数据是一致的。引擎抓取页面信息时，需要做一定的重复内容检测，遇到内容大量抄袭、重复采集的网站，可以不再爬取。数据存储方式多样，可以存入本地数据库也可以存入临时数据库,还可以存入 txt 文件或 csv 文件。

2、数据清洗

爬取到的数据通常会有些杂乱，关于 python 数据处理过程中的数据预处理，是主要对缺失值/空格/重复值的数据清洗。对采集到的数据文件或数据库进行文件读写操作，对数据表处理，再运用预处理函数对数据的重复值、异常值、缺失值进行处理。

3、数据可视化处理

（1）确定问题，选择图形

依据课题研究对象确定模型，从直方图、柱状图、折线图、散点图、饼图、气泡图、雷达图、地图等众多图标类型中选择能恰当描述、解释问题的图形。

（2）转换数据，应用函数

将数据清洗步骤中整理好的数据运用到可视化处理中，根据第一步选择好的图形，去找 python 中对应的函数。

（3）参数设置，一目了然

根据需求修改颜色（color），线型（linestyle），标记（maker）或者其他图表装饰项标题（Title），轴标签（x，y），轴刻度（set_xticks），还有图例（legend）等，

4、预测型数据分析

对于需要预测的不同变量进行分类，采用监督学习算法，对于目标变量是离散型类别，可以看作分类问题，采取 Logistic 模型；对于目标变量是连续型类别，可以看作回归问题，采用线性回归。如果样本既包含离散型变量又包含连续型变量，可以根据样本具体特征分析选择 K 近邻算法，决策树算法或者随机森林算法

六、预期成果形式描述

1、二手房数据分析预测系统模型代码

2、毕业论文

七、指导教师意见

同意开题

指导教师签名：魏艺

2022 年 12 月 25 日

八、学院（部）意见

同意

学院（部）（盖章）

2022 年 12 月 25 日

山东师范大学

毕业论文教师指导记录表

<table>
<tr><td>学院（部）</td><td>信息科学与工程学院</td><td>专业</td><td>计算机科学与技术（公费师范生）</td><td>班级</td><td></td></tr>
<tr><td>姓名</td><td>常子儒</td><td>学号</td><td colspan="2">201911990102</td><td>指导教师</td><td>魏艺</td></tr>
<tr><td>第一次指导</td><td colspan="6">选题及开题报告指导
通过QQ软件的方式与老师讨论了论文大致的研究方向与开题报告需要完成的内容，老师帮助给出了开题报告国内外研究现状的撰写示例，指明需要阅读的参考文献类型等，并对开题报告中各部分的格式做了要求。在老师帮助下，论文开题报告按要求完成提交</td></tr>
<tr><td>第二次指导</td><td colspan="6">论文系统开发任务指导
通过QQ软件的方式老师询问了论文进度及所遇到的困难，与老师讨论了二手房数据分析预测系统所需数据集的获取与系统实现功能，老师对爬虫爬取的数据来源、可能困难与数据集大小等方面作了详细的指导并说明了毕业设计最终实现结果。</td></tr>
<tr><td>第三次指导</td><td colspan="6">论文撰写指导
在将论文初稿交给老师审阅后，老师给出了论文初次修改意见，指出论文的系统实现部分应该增添关于爬虫的内容，具体问题像如何实现的爬虫、从哪里爬取的数据、数据量有多少等等。另外老师对内容篇幅提出要求，需要将论文中部分概念或技术介绍等做简化。</td></tr>
<tr><td>第四次指导</td><td colspan="6">论文撰写报告
通过QQ软件的方式，与老师讨论了系统实现的数据获取部分，老师指出了系统所需数据集需要给出具体的数据来源，另外论文篇幅较长，所用图表较多，老师对论文中图表、表格与具体段落文字格式等给出了修改意见，使论文更加严谨和美观。最后，老师对论文最终版进行了修</td></tr>
</table>

	改与检查，包括中英文摘要、论文数据表格与错别字等。 通过

山东师范大学本科生毕业论文（设计）

指导教师意见

<table>
<tr><td>学院（部）</td><td>信息科学与工程学院</td><td>专业</td><td>计算机科学与技术（公费师范生）</td><td>班级</td><td></td></tr>
<tr><td>姓名</td><td>常子儒</td><td>学号</td><td colspan="2">201911990102</td><td>指导教师</td><td>魏艺</td></tr>
<tr><td colspan="2">论文（设计）题目</td><td colspan="5">基于 Python 的二手房数据分析预测系统</td></tr>
<tr><td colspan="7">指导意见：
论文选题符合计算机专业培养目标的要求，以潍坊市二手房交易市场为研究对象，设计并开发了一个基于 python 的二手房数据分析预测系统。该系统提供了数据查询、数据图表展示及房价预测等功能，使用户能深入挖掘二手房数据信息并帮助其做出决策。论文内容完整，语言流畅、层次结构安排科学，对问题的洞察比较透彻，相关分析论证有一定参考价值，表明该同学对相关的专业知识掌握较好，具有一定的分析问题和解决问题的能力。本论文达到了学士学位应有的水平，同意其参加论文答辩。</td></tr>
<tr><td>成绩</td><td>98.0</td><td>指导教师</td><td>魏艺</td><td colspan="3">2023 年 4 月 29 日</td></tr>
</table>

山东师范大学本科生毕业论文（设计）

评阅人意见

学院（部）	信息科学与工程学院	专业	计算机科学与技术（公费师范生）	班级	
姓名	常子儒	学号	201911990102	指导教师	魏艺
论文（设计）题目	基于 Python 的二手房数据分析预测系统				
评阅人意见： 论文通过爬虫从二手房交易网站上收集了潍坊市的二手房交易信息，设计并实现了一个基于 python 的二手房数据分析预测系统，该系统提供了数据查询、数据图表展示、房价预测等功能，可以帮助用户深入挖掘二手房数据信息并做出决策。论文选题符合专业培养目标，能够达到综合训练目的，题目有一定难度，工作量较大，选题具有较大的实践意义。论文内容完整，层次结构清晰，逻辑性较强，表明该同学具备了一定的独立工作能力。论文文字流畅，格式符合学位论文规范要求，达到了学士学位论文专业水平要求。					
成绩	96.0	评阅人	张宇昂	2023 年 5 月 15 日	

山东师范大学本科生毕业论文（设计）

答辩委员会意见

<table>
<tr><td>学院（部）</td><td>信息科学与工程学院</td><td>专业</td><td>计算机科学与技术（公费师范生）</td><td>班级</td><td></td></tr>
<tr><td>姓名</td><td>常子儒</td><td>学号</td><td colspan="2">201911990102</td><td>指导教师</td><td>魏艺</td></tr>
<tr><td>论文（设计）题目</td><td colspan="6">基于 Python 的二手房数据分析预测系统</td></tr>
<tr><td colspan="7">答辩委员会意见：
该同学的论文选题具有理论与现实意义，文献材料收集详实，综合运用了所学知识，通过搭建潍坊市二手房数据分析预测网站来使用户深入挖掘二手房数据信息并帮助其做出决策，有一定的应用价值。答辩过程中，该同学能够条理清晰地将设计过程讲述出来并有理有据地对提问问题作答，具有较强的表达能力和学习能力。综合指导教师、评阅人意见以及学生答辩过程表现，经答辩委员会认真讨论，一致同意通过该同学的毕业论文答辩，同意授予学士学位。</td></tr>
<tr><td>成绩</td><td>95.0</td><td>答辩主席</td><td colspan="2">张宇昂</td><td colspan="2">2023 年 5 月 16 日</td></tr>
</table>

山东师范大学

毕业论文答辩记录表

学院（部）	信息科学与工程学院	专业	计算机科学与技术（公费师范生）	班级	
姓名	常子儒	学号	201911990102	指导教师	魏艺
论文（设计）题目	基于 Python 的二手房数据分析预测系统				

答辩记录：

答辩时间：2023 年 5 月 16 日 14：00

答辩地点：文淙楼 406

评委老师：张宇昂，嵇存，魏艺，边际，张梦洋

答辩题目：基于 Python 的二手房数据分析预测系统

提问与回答：

问题一：简要介绍下你论文所做的工作。

答：论文实现了二手房数据分析系统网站，从数据获取、系统搭建与测试部署三部分入手，设计了个人信息记录、数据查询、数据可视化、房价预测四个模块，旨在为用户提供数据分析功能与决策支持。

问题二：你的论文中部分英文与数字没有使用 Times New Roman 字体。

答：抱歉，老师。论文所用到的英文与数字较多，部分英文与数字在论文写作时默认使用了宋体，对于涉及到的内容，我会尽快修改的。

问题三：论文所展示的系统网站图片模糊。

答：抱歉，老师。由于网站页面内容较多，为保证完整性，就截取了全部内容。在插入图片后选择了缩小，使得图片模糊。我会选择合适的图片替换模糊图片。

问题四：表格字体与间距看起来比较大。

答：抱歉，老师。论文表格选择了与正文同一字号与间距，以上错误我会尽快修改，将字号与间距分别调成标准格式。

答辩录入员	张梦洋	2023 年 5 月 16 日

山东师范大学本科毕业论文（设计）摘要

学院： 信息科学与工程学院 专业： 计算机科学与技术（公费师范生） 班级：

<table>
<tr><td>姓名</td><td>常子儒</td><td>学号</td><td>201911990102</td><td>指导教师</td><td>魏艺</td></tr>
<tr><td colspan="2">论文（设计）题目</td><td colspan="4">基于 Python 的二手房数据分析预测系统</td></tr>
<tr><td colspan="2">关键词</td><td colspan="2">二手房交易；Flask 框架；数据分析；数据可视化；线性回归</td><td>论文（设计）字数</td><td>24790</td></tr>
<tr><td colspan="6">内容摘要：
本文以潍坊市二手房交易市场为研究对象开发了基于 python 的二手房数据分析预测系统，旨在为房地产行业提供数据、图像支持和决策依据。
本文所做工作总共分为数据获取、系统搭建与测试部署三部分，在数据获取部分中利用 requests、lxml 等第三方库从链家等二手房交易网站上爬取潍坊市二手房信息并存储，并解决了在长时间、大批量数据获取时的连接中断与请求失败的问题；在系统搭建部分中通过 Flask 框架确定系统网站逻辑，并将获得的数据信息进行清理并作预处理，然后利用 numpy、pandas、matplotlib 等数据分析库对预处理完成的数据作可视化来了解数据的特征与分布，最后利用 Scikit-learn 机器学习库建立多元线性回归模型与 K 近邻模型并训练优化，利用交叉验证来评估模型性能；在测试部署部分中分别在不同环境下对系统网站进行功能性测试与非功能性测试，同时将系统网站部署到 ECS 云服务器中实现公网访问。
本文建立的二手房数据分析预测系统提供数据查询、数据图表展示及房价预测等功能，使用户能深入挖掘二手房数据信息并帮助其做出决策。</td></tr>
<tr><td>成绩</td><td>97.0</td><td colspan="2">（学院公章）</td><td colspan="2">2023 年 5 月 30 日</td></tr>
</table>