

**《Python语言数据分析与挖掘》**

**基于机器学习模型建立回归分析模型进行房价预测及济南二手房爬虫、整体情况分析**

|  |  |
| --- | --- |
| 学 院： | 信息工程学院 |
| 专 业： |  |
| 班 级： |  |
| 组 号： | 3 |
| 组长签字： |  |
| 成绩评定： |  |

信息工程学院制

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年级 | 2021 | | 专业 | | 计算机科学与技术 | 班级 | 2 |
| 组号 | 3 | | 队名 | | 金铲铲队 | | |
| 队呼 | 开拓创新，勇往直前 | | | | | | |
| 主题 | 基于机器学习模型建立回归分析模型进行房价预测及济南二手房爬虫、整体情况分析 | | | | | | |
| 小 组 成 员 | | | | | | | |
| 姓名 | | 学号 | | 组内分工 | | | |
|  | |  | | 数据集的爬取，资料收集 | | | |
|  | |  | | 数据的预处理，数据清洗 | | | |
|  | |  | | 数据模型建立，情景测试 | | | |
|  | |  | | 数据分析及可视化，文档整理 | | | |
|  | |  | |  | | | |
|  | |  | |  | | | |
| 团队照片 | | | | | | | |

1. **项目背景**

房子是人们生活中必不可少的东西。现在的年轻人越来越理智了，对于爱情和面包的分量还是能掂量清的，尤其是在谈到结婚问题时，都会优先考虑到房子的问题。我们几年后就毕业工作，有了结婚需求，买房问题自然提上日程。另一方面，由于工作的需要，我们毕业后工作，随之带来的是租房的需求。而福利水平和租房生存所必须的配套建设都还不完善，租房注定是权宜之计。目前很多地方对于申请廉租房都有严格的条件限制，加上房源数量有限，不是所有的大学生一毕业就可以住上“国家提供的房子”。所以毕业入手二手房是不错的选择。

1. 选择面更广，价格合理。二手房价格区间大，可供选择的余地多。这对于毕业生来说压力是远远小于新房的。

2. 交通更便利，上班方便。一般来说，二手房由于建设的年份稍久，所以距离市中心往往比较近。这样方便了上下班，对于平时生活也是诸多便利。

还有很多优点，比如周围配套设施比较完善，休闲娱乐比较方便，节省交通费，产权权属清晰，入住时间快等待时间少等。

我们的项目主要通过爬取二手房信息，对爬取的数据进行进一步清洗，分析各维度的数据，筛选对房价有显著影响的特征变量，探索济南二手房整体情况、价格情况和价格的影响因素，建立房价预测模型。主要流程包括：

1. 数据的爬取收集
2. 数据的预处理，数据清洗
3. 数据分析及可视化
4. 建立房价预测模型，模型测试
5. **问题分析**

* 探究单价、数量、总价和行政区域之间的关系
* 探究其它因素和总价的关系
* 户型分布
* 分析建筑年代情况
* 使用机器学习模型建立回归分析模型进行价格预测

1. **数据获取方法及过程数据预处理（附主要代码）**
2. 数据获取

通过爬虫，爬取了链家网官网济南市全市共11个行政区的详细的二手房房屋数据供50390条。

数据主要内容包括：小区名称、行政区、区域、参考总价、参考单价、房屋户型、所在楼层、建筑面积、户型结构、套内面积、建筑类型、房屋朝向、建筑结构、装修情况、梯户比例、配备电梯、挂牌时间、交易权属、上次交易时间、房屋用途、房屋年限、产权所属、抵押信息、房本信息、房协编号、经度、纬度等全部可获取到的详细数据。

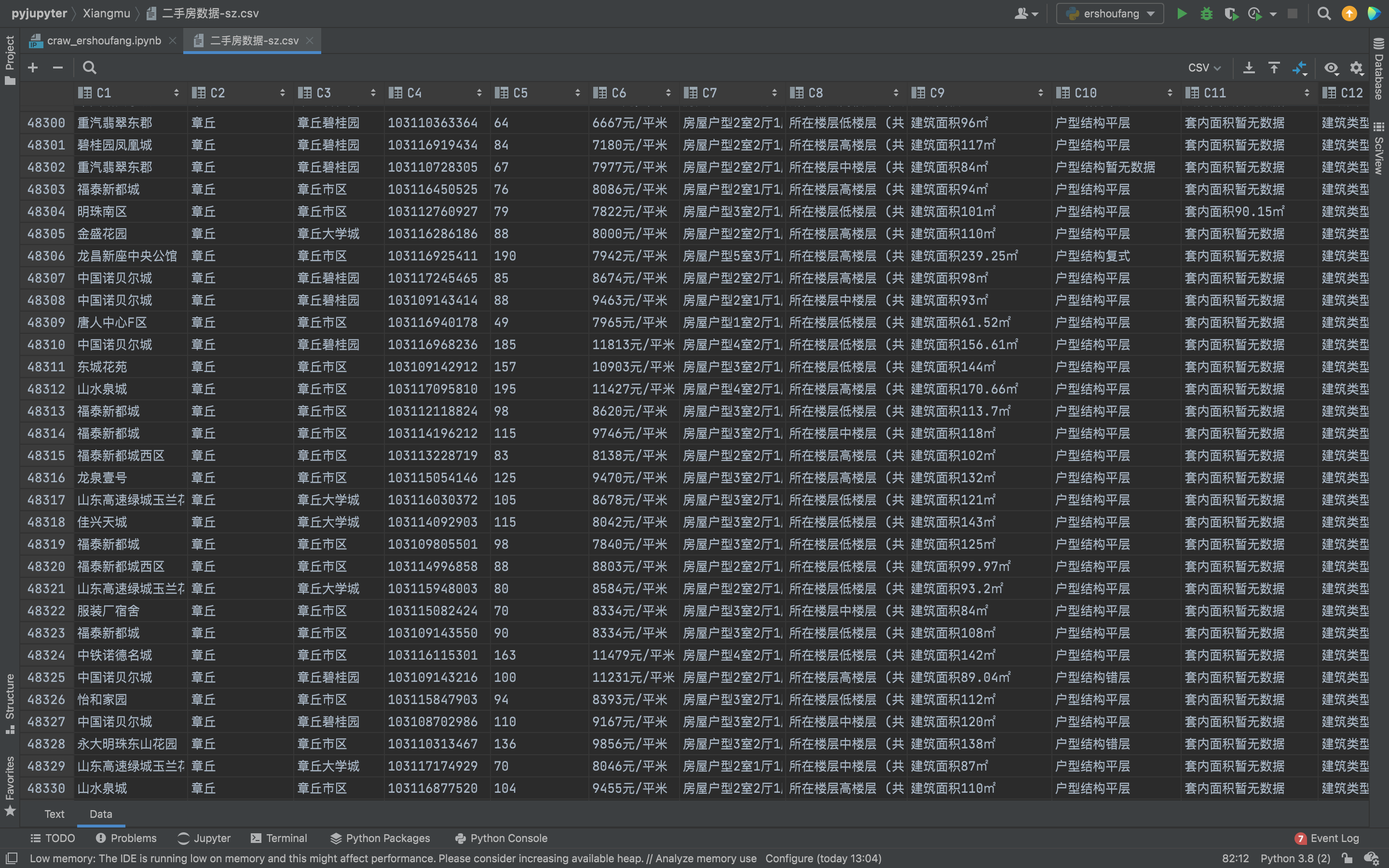


图 3-1 数据概览

1. 数据预处理

获取到详细的房屋数据后，对所有的数据进行数据预处理并进行简单的数据分析，充分的了解所获取的房屋数据的整体情况，得到我们想要使用的数据集。

1.各行政区域二手房数量

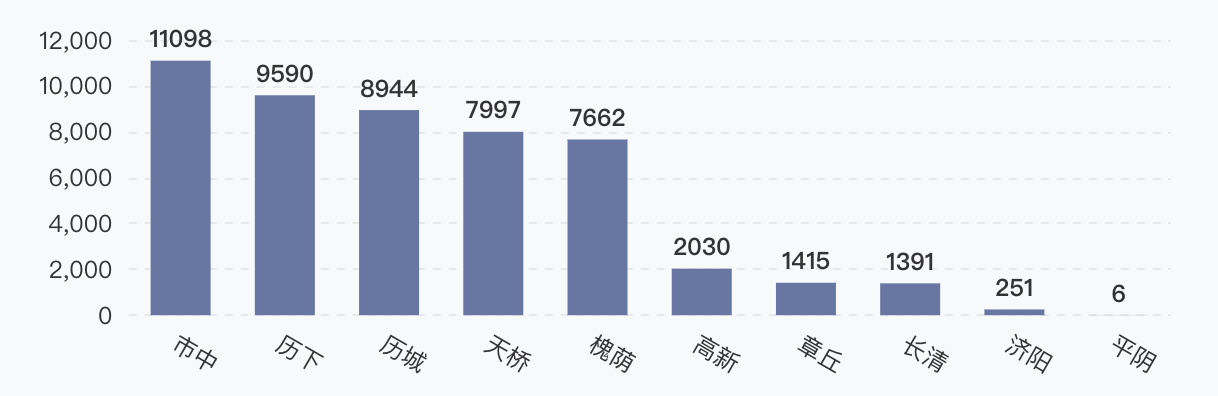


图 3-2 各行政区域二手房数量

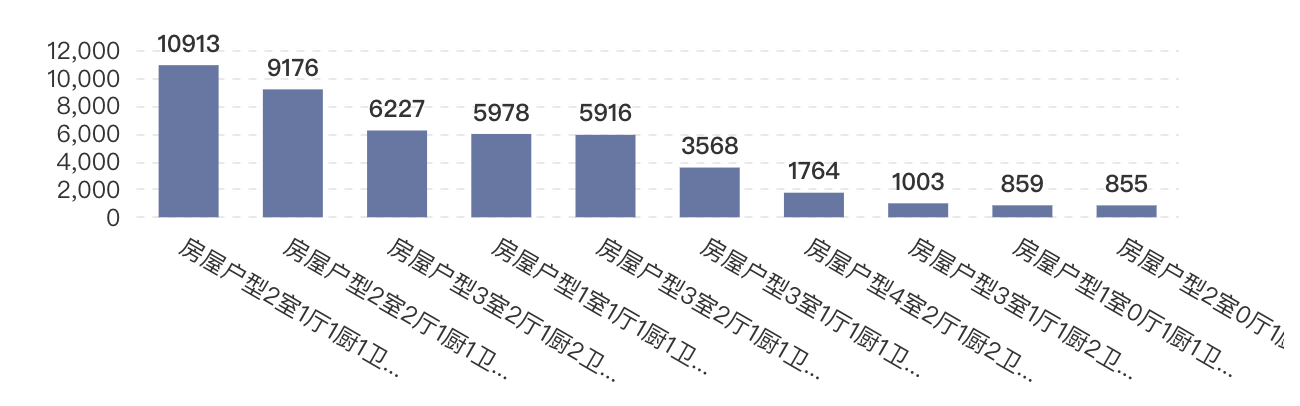
2.各居室类型的房屋数量

图 3-3 各居室类型的房屋数量

3.各种楼层分布大致情况

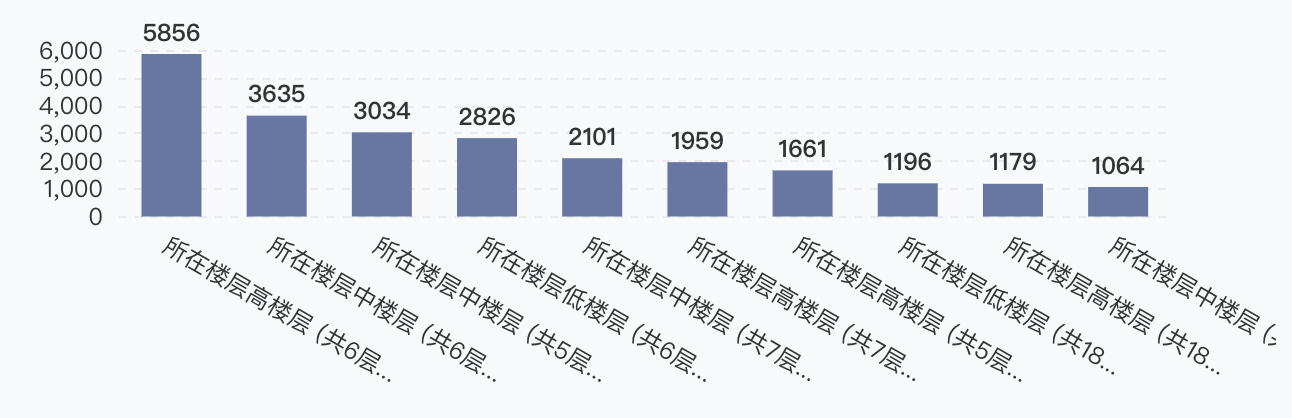


图 3-4 各种楼层分布大致情况

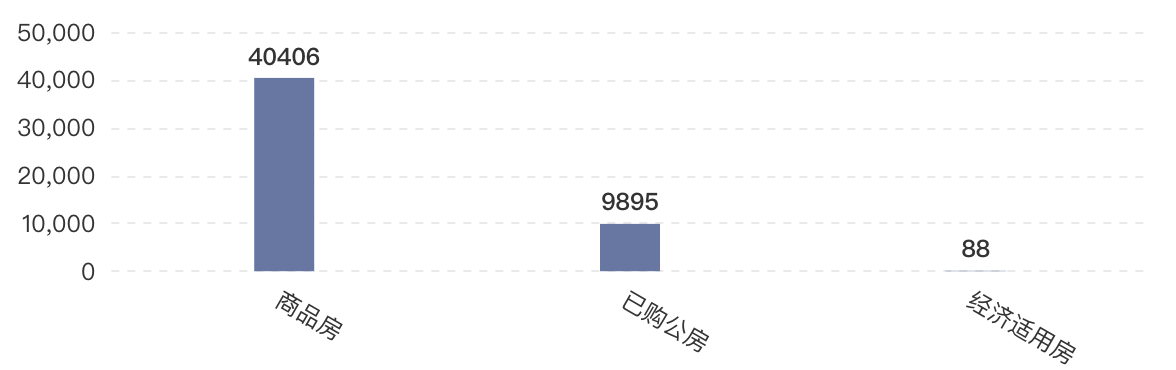
4.各种房屋产权属性情况

图 3-5 各种房屋产权属性情况

5.各种房屋用途的统计



图 3-6 各种房屋用途的统计

1. **数据处理**

1.数据预处理

1. 导入数据集，查看数据整体情况



图 4-1 查看数据集整体情况

1. 判断是否有重复项/缺失值

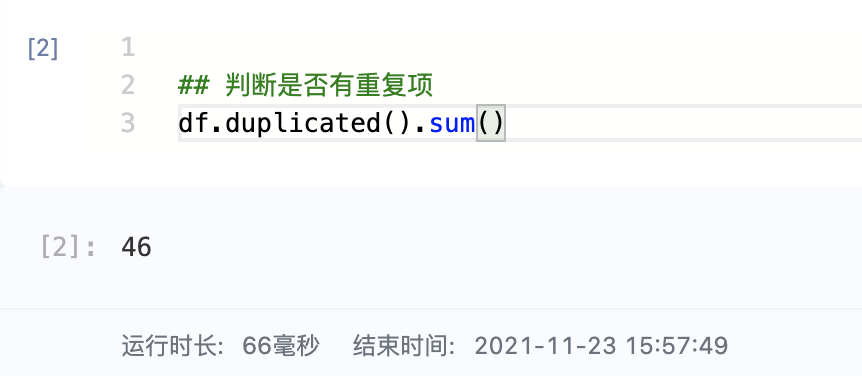


图 4-2 重复项判断



图 4-3 判断缺失值

1. 查看数据类型



图 4-4 查看数据类型

1. 唯一标签值



图 4-5 唯一标签值

* 对数据去重、缺失值处理
* 建筑面积、年代、单价需要进行转换（去掉单位）
* 楼层、区域等数据整合

2.数据清洗

1. 数据格式转换
2. 重复值处理
3. 缺失值处理

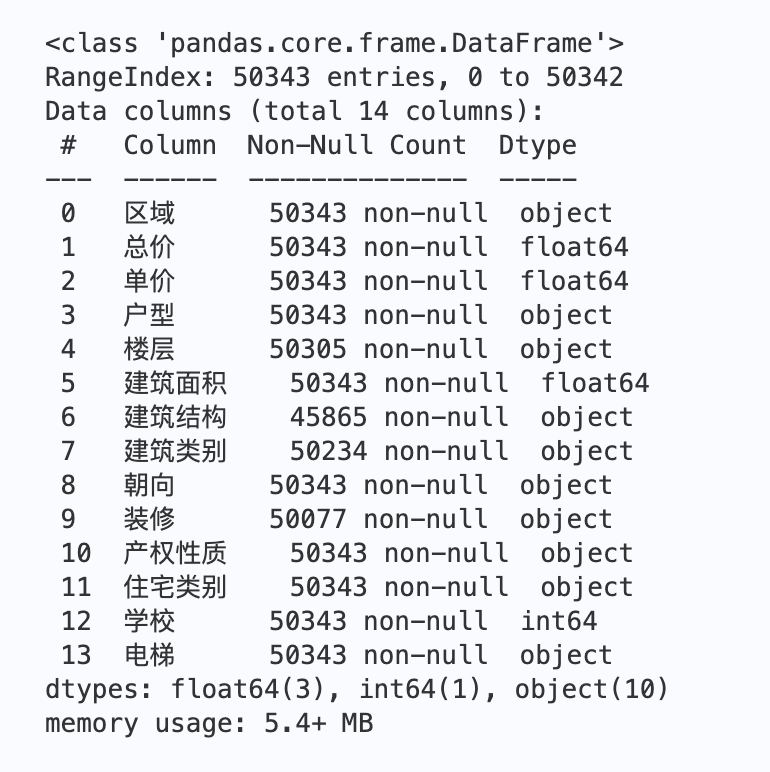


图 4-6 格式转换/重复值处理/缺失值处理结果

1. 异常值处理

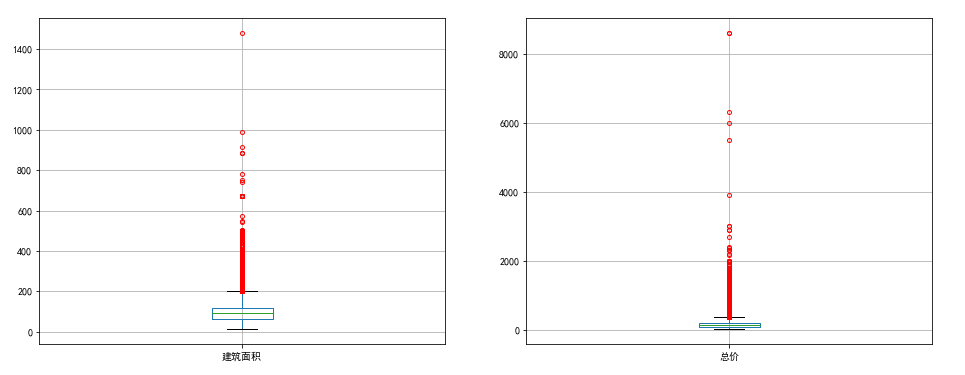


图 4-7 异常值处理箱线图分析

1. 描述性分析



图 4-8 描述性分析数据

* 二手房最小面积为10.8平米，最大面积为1481平米
* 最便宜的15万，最贵的8600万。
* 面积大概集中在64-117.91平米，价格大概集中在94-210万。
* 将总价高于上限的当作异常值进行处理（直接删除法），考虑大多数人可购买的情况。

1. **数据可视化**
2. 探究单价，数量，总价和行政区域之间的关系

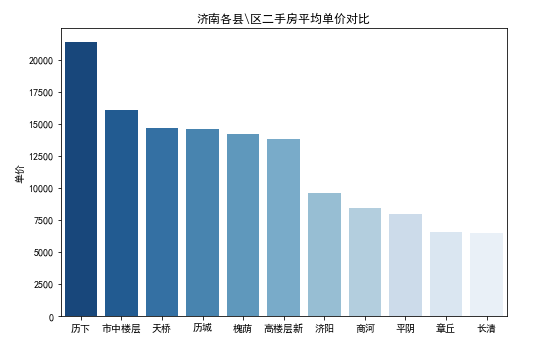


图 5-1 济南各县/区二手房平均单价对比

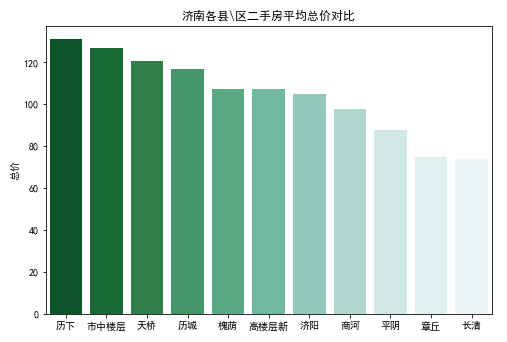


图 5-2 济南各县/区二手房平均总价对比

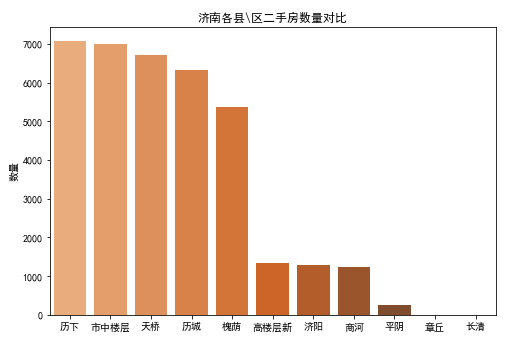


图 5-3 济南各县/区二手房数量对比

* 济南各县区二手房平均单价、总价、数量都是一样的排列顺序，由高到低，最高的是历下区，最低的是长清区。

1. 探究面积和总价的关系

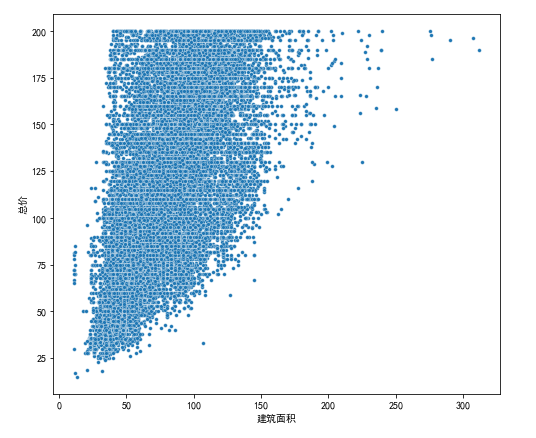


图 5-4 朝向和总价的关系

基本服从面积越大，价格越高的关系。

1. 探究朝向和总价的关系

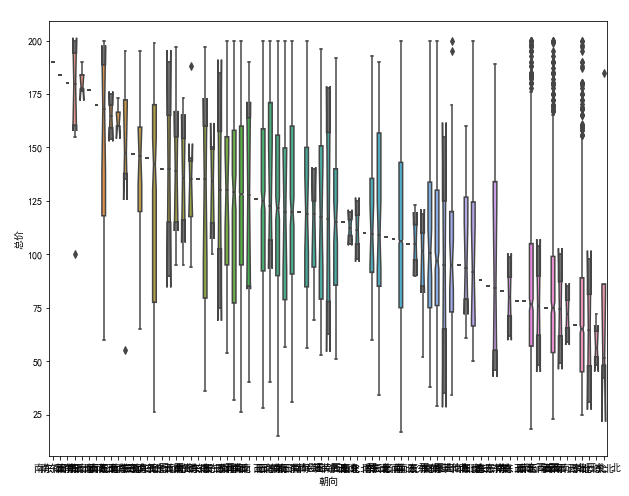


图 5-5 朝向与总价的关系

* 风水：人们在挑选房子时，经常喜欢挑坐北朝南的房子，因为这种房子采光好，顺光顺水，冬暖夏凉，很适合人居住。
* 包含南、北朝向方位的价格相对来说要贵一点。

1. 探究装修和总价的关系

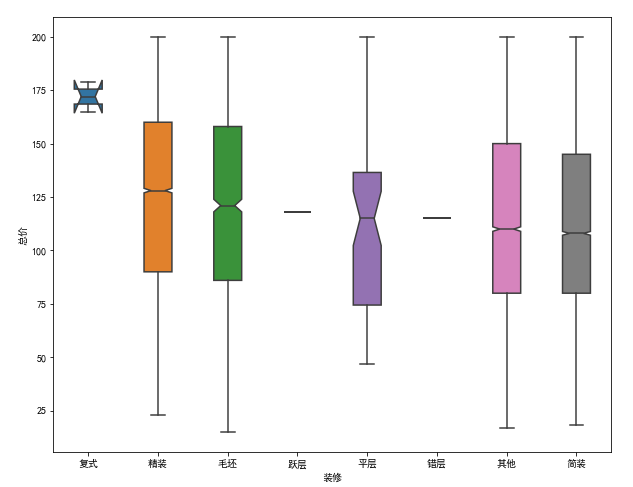


图 5-6 探究装修和总价的关系

* 不同装修信息对总价是有一定影响的，装修情况越好价格会偏高一点。

1. 探究楼层和总价的关系

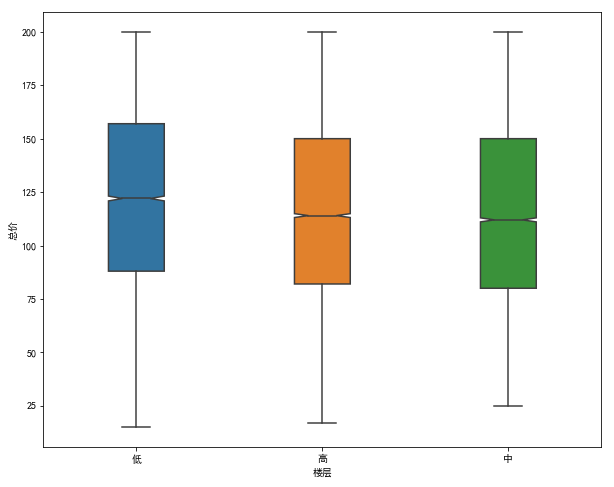


图 5-7 探究楼层和总价的关系

* 不同楼层对总价影响较小。

1. 探究电梯和总价的关系

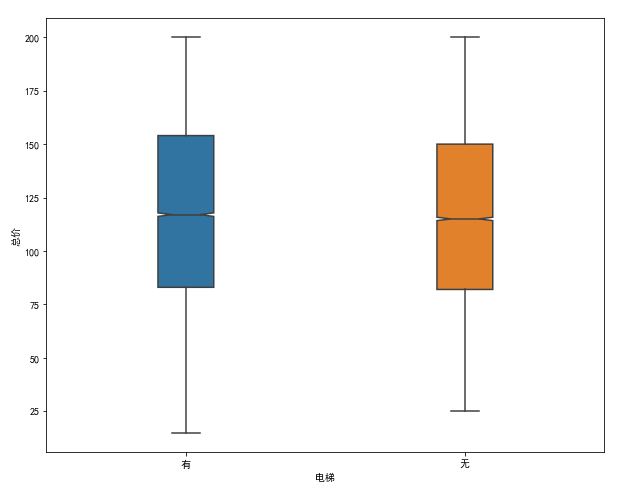


图 5-8 探究电梯和总价的关系

* 有电梯的房子比没电梯的房子要贵。

1. 探究学区房和总价的关系

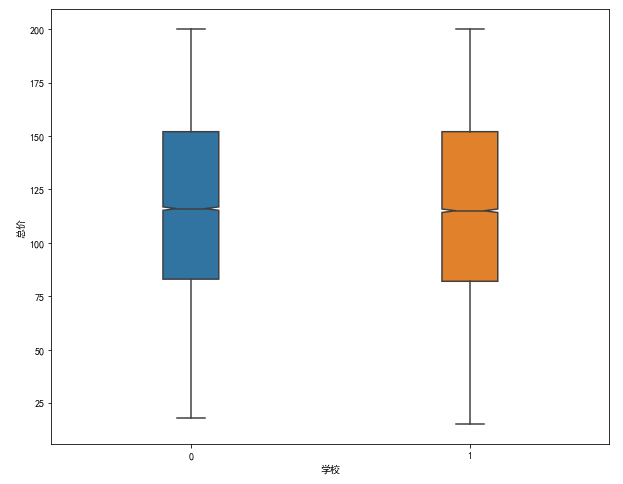


图 5-9 探究学区房与总价的关系

* 附近有学校的房子价格会高一些。

1. 建筑年代情况分析以及和总价的关系

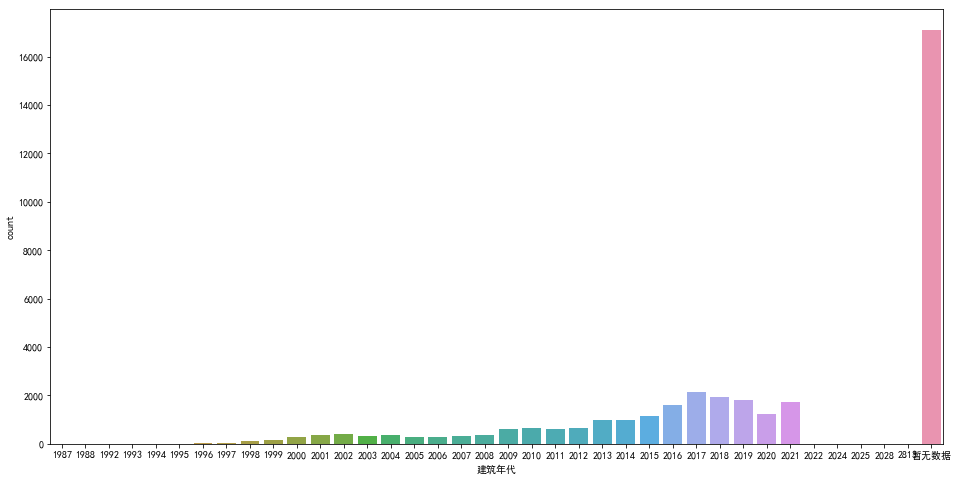


图 5-10 建筑年代情况分析

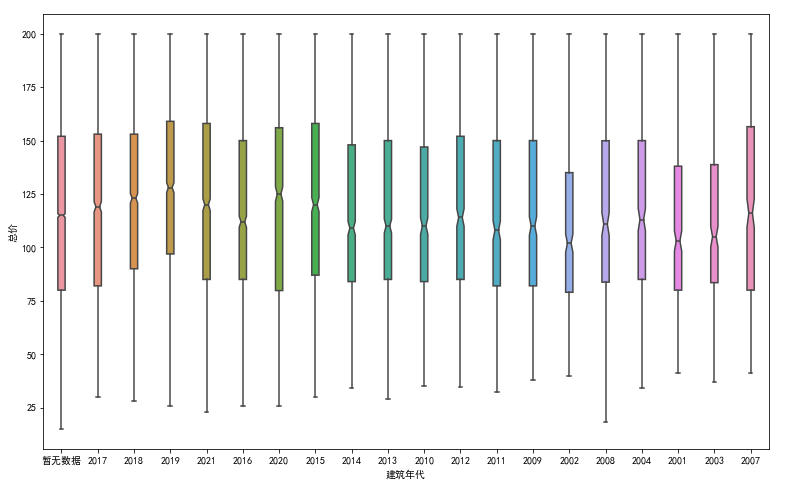


图 5-11 建筑年代与总价的关系

* 出售的二手房大都是几年前的，比较符合现实情况，新房很少会有人出售的。
* 年代太久的房子价格会偏低，在2008年之后的房子价格会比较高，最新的房子（2017年之后）价格不如之前年代的。

1. 产权性质、住宅类别、建筑结构、建筑类别与总价的关系

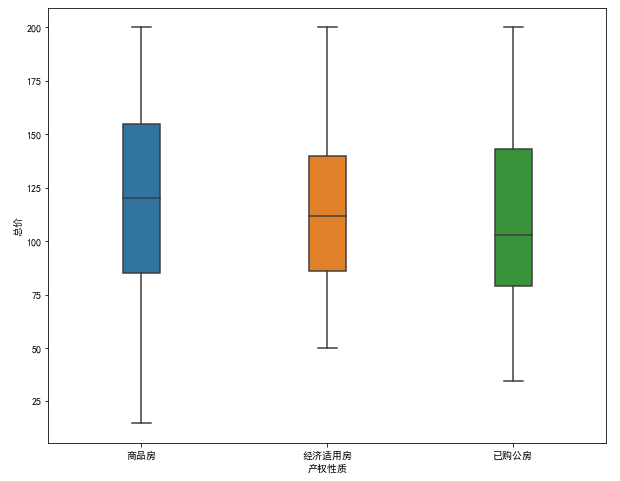


图 5-12 产权性质与总价的关系

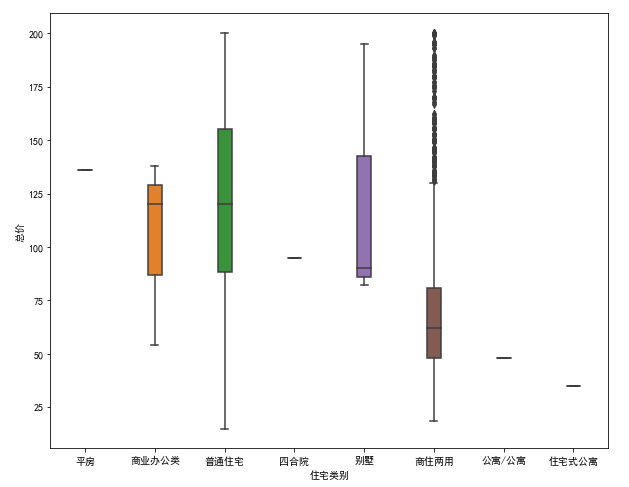


图 5-13 住宅类别与总价的关系

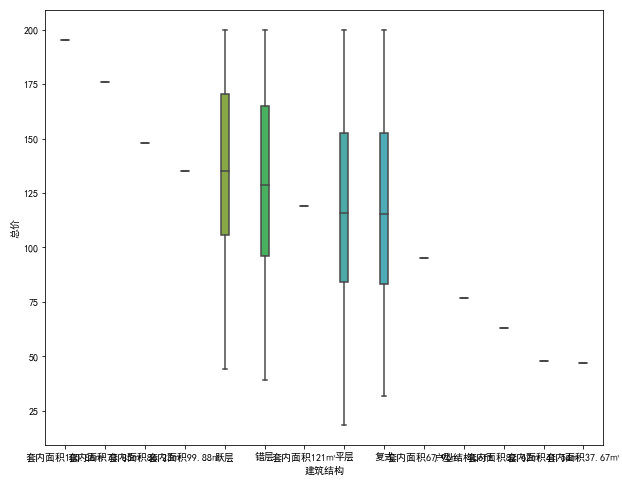


图 5-14 建筑结构与总价的关系

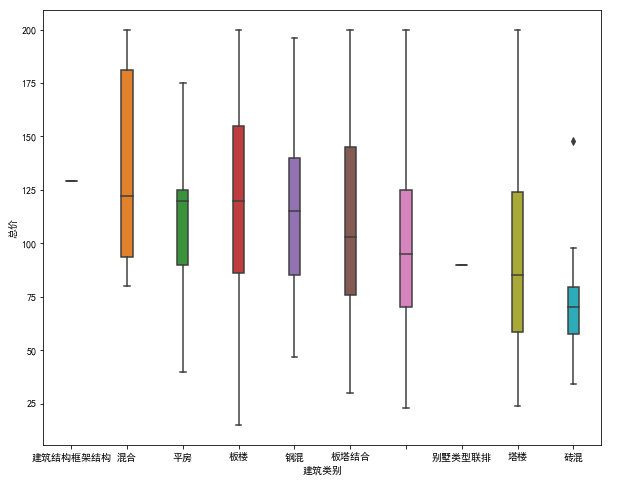


图 5-15 建筑类别与总价的关系

* 这几个变量对价格都有一定的影响力。

1. 户型和总价的关系

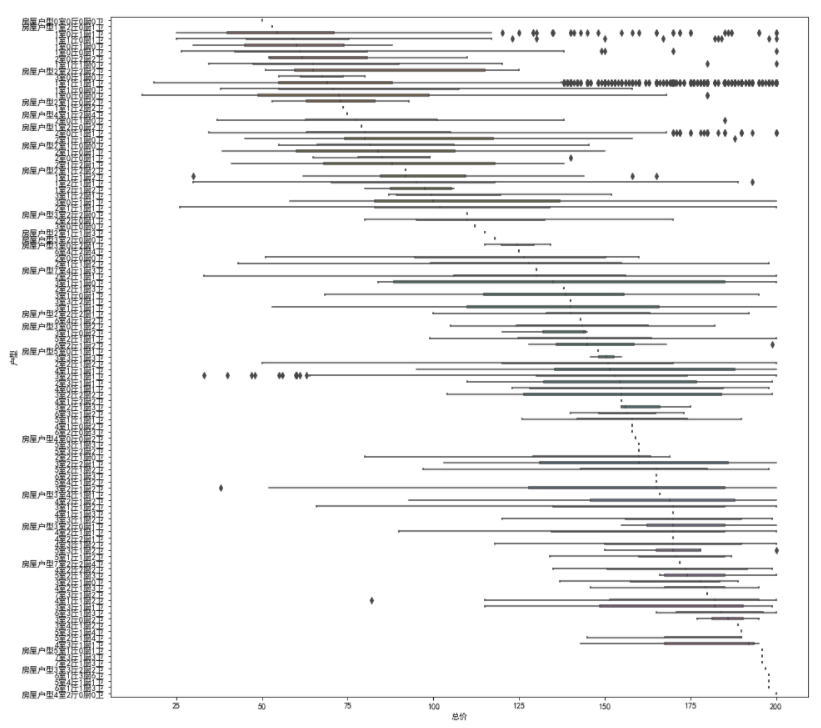


图 5-16 户型与总价的关系1

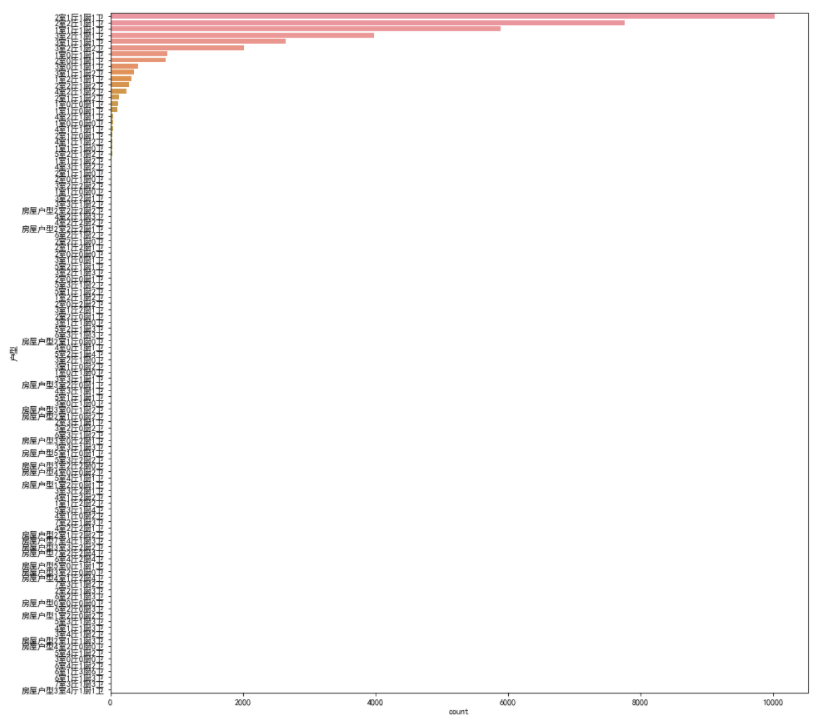


图 5-17 户型与总价的关系2

* 户型对总价影响比较大，不同的室、厅、卫又会产生不同的影响。
* 大多数需求集中在 2或3 室 1或2 厅 1或2 卫。

1. **数据建模、模型评估**
2. 删除所有缺失值
3. 分解户型
4. 删除楼层、户型、单价
5. 编码-有序多分类（根据上面可视化结果，按照对价格的影响程度排序，越大影响越高）无序多分类无法直接引入，必须“哑元”化变量等级变量（有序多分类）可以直接引入模型
6. 删除X年的房子，年代转变为房龄
7. 套索回归

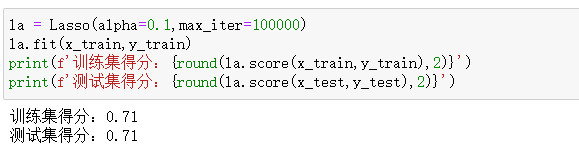


图 6-1 套索回归结果

1. 随机森林

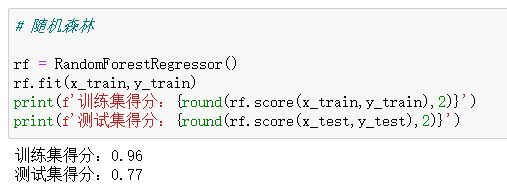


图 6-2 随机森林结果

1. 决策树

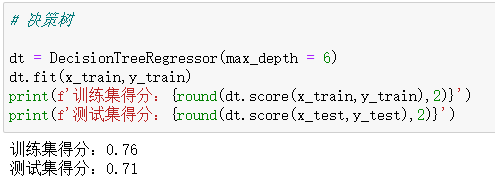


图 6-3 决策树结果

1. K近邻

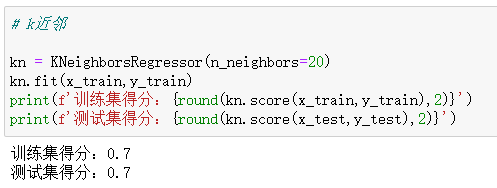


图 6-4 K近邻结果

* 比较几种模型，最终在测试集的得分都能保持在70%以上。
* 随机森林训练集得分达到90%以上，测试集得分在几种模型里表现也是最好的。

**模型测试：**

一家三口，孩子即将上学，大人历下区工作，需要购买房子，假设要求如下：

* 3室1厅1卫（3、1、1） 面积大概再95㎡左右（95）
* 学区房（1） 东南（10）
* 中装修 （4） 无电梯 （0）
* 个人产权（5） 普通住宅（4）
* 平层（4） 钢混（5）
* 城关（6） 房龄 （10）

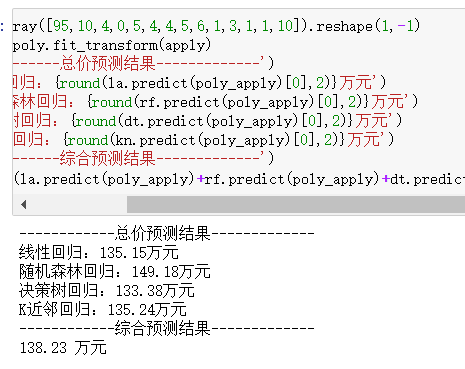


图 6-5 通过模型进行房价预测

1. **结论**

通过爬取济南市当地最新二手房信息，对爬取的数据进行进一步清洗，分析各维度的数据，筛选对房价有显著影响的特征变量，探索济南二手房整体情况、价格情况和价格的影响因素，建立房价预测模型。

 1. 二手房最小面积为10.8平米，最大面积为1481平米

 2. 最便宜的15万，最贵的8600万。

 3. 面积大概集中在64-117.91平米，价格大概集中在94-210万。

 4. 济南各县区二手房平均单价、总价、数量都是一样的排列顺序，由高到低，最高的是历下区，最低的是长清区。

5．基本服从面积越大，价格越高的关系。

 6. 人们在挑选房子时，经常喜欢挑坐北朝南的房子，因为这种房子采光好，顺光顺水，冬暖夏凉，很适合人居住，大多数人还是比较注重风水的。

 7. 包含南、北朝向方位的价格相对来说要贵一点。

 8. 不同装修信息对总价是有一定影响的，装修情况越好价格会偏高一点。

 9. 不同楼层对总价影响较小。

 11. 有电梯的房子比没电梯的房子要贵。

 12. 附近有学校的房子价格会高一些

 13. 出售的二手房大都是几年前的，比较符合现实情况，新房很少会有人出售的。

 14. 年代太久的房子价格会偏低，在2008年之后的房子价格会比较高，最新的房子（2017年之后）价格不如之前年代的。

 15. 产权性质、住宅类别、建筑结构、建筑类别这几个变量对价格都有一定的影响力。

 16. 户型对总价影响比较大，不同的室、厅、卫房型又会产生不同的影响。

 17. 大多数需求集中在 2或3 室 1或2 厅 1或2 卫。

18. 通过房价预测模型，可以得出较好的预测结果，切合实际。

1. **遇到问题及解决办法**

1. 数据爬取过程第一步，由于需求数据项比较多，如何设置恰当的筛选条件是首要问题。通过网络搜索研究各种爬取案例，最终实现了数据项的多次筛选。

2. 数据爬取过程涉及详情页的翻页，页面信息的参数等。通过F12查看源码了解具体的页面布局以及各个数据项的关键字完成爬虫参数设置。

3. 数据爬取过程中解析内容可能是HTML，可能是JSON,可能是二进制。在这里解析内容是HTML，我们通过使用正则表达式，网页解析库进行解析。

4. 数据处理过程，由于数据爬取完比较乱，在处理过程中，通过学习Excel知识点完成数据的处理。

5. Jupyter notebook无法弹出窗口绘图。解决方法：弹出窗显示必须加上：%matplotlib auto 语句。

6. 使用matplotlib显示图片时，中文字体显示问题报错。网络查找原因发现matplotlib默认不支持中文，通过查找字体库的存放路径，下载新的中文字体，放到字体库文件中，解决报错。

1. **收获**

**邹世昕：**

爬虫，首先，python3的语法必须知道，不过好在语法也非常简洁。但是，写着发现有很多问题，就是比如一个形参，由于不确定类型，.无法像java那样补全，导致如果忘了一个函数，必须查看代码或手册，对于基础的库的学习深度也不够，返回的数据也是各种各样的类型，如何选择合适的解析器，这些只能边查边用。在这个过程中碰到的任何问题，都要很有耐心的去解决，通过查阅浏览器或者和他人讨论，来解决问题

**韩章豪：**

在整个项目的数据处理清洗过程总，用到的主要也是已经学习了解的numpy、pandas和数据挖掘的知识，由于数据比较多，对于数据的整体分析和理解还是比较有难度的，处理过程虽然不是很难，但是却需要很大耐心去处理，通过项目的运用，加深了对已经学过的相关知识，也了解代理数据挖掘过程中的复杂性和难度，想要学好它，还需要花费很长的时间。

**杜宝敏：**

我对于数据可视化了解并不深入，跟大多数人一样，直观地认为是各种统计图形的使用和设计。在项目中也只是简单的运用matplotlib完成了图形的显示。但是若有更深入的需求，那么自主设计与开发主题图才更加重要。这意味着不仅要有足够的设计能力，还的要有足够的编码能力和软件体系架构的能力。最后有两点很重要：第一，可视化设计不等于不同统计图那么简单；第二，要针对具体业务进行可视化设计，少不了先做好对数据的前期清洗工作。

**翟肇雨：**

通过此次项目，和小组成员明确分工，积极完成各自的任务。我在其中，主要负责数据建模和模型的情景测试，在完成项目过程总，对数据挖掘和建模有了更多了解，如何根据对象特征和目的，对问题进行简化完成模型的建立，如何验证模型的合理性达到符合实际应用。虽然只是简单的接触使用，但是这个过程对于逻辑能力分析能力有很好的锻炼和提高，带给我更多的发散性思维。