**数据仓库与数据挖掘项目报告**

梅浩楠，徐尘化

2020.12.6

目录

[一． 项目目的及意义 1](#_Toc58103317)

[二． 技术路线 1](#_Toc58103318)

[三． 研发情况 2](#_Toc58103319)

[1. 数据预处理 2](#_Toc58103320)

[2. 可视化分析 3](#_Toc58103321)

[3. 价格预测及聚类分析 8](#_Toc58103322)

[4. 可视化页面 8](#_Toc58103323)

[5. 卖方价格预测页面 10](#_Toc58103324)

[6. 买方房屋推荐页面 10](#_Toc58103325)

[四． 成果总结 10](#_Toc58103326)

[1. 完成情况 10](#_Toc58103327)

[2. 人员分工 10](#_Toc58103328)

[五． 项目展望 11](#_Toc58103329)

# 项目目的及意义

项目的目的是在采集自贝壳二手房交易平台的成都市二手房成交数据的基础上，对数据进行处理和挖掘，以网站为载体实现二手房交易分析、卖方价格预测和买方房屋推荐三个主要功能。

二手房交易分析功能服务于统计人员，对成都的二手房交易做完整的统计分析，使用折线图、散点图、饼图等展示属性间的关系，提供交互功能，可以对成都市行政区划地图进行点选，展示不同区的具体信息并选择不同的区进行对比。

卖方价格预测功能服务于卖方，通过用户提供的二手房的各种属性预测其成交价，为二手房交易中的卖方提供一个参考值，使卖方对于自己房子的价格有一个较为准确的估计，并在地图上标记房子。

买方房屋推荐功能服务于买方，通过买方选取的各个属性可接受的范围，为买方列出可以选择的二手房范围，在地图上进行标记，使买方在购房前有一个心理预期。

# 技术路线

这是预计的技术路线图：



以下为实际技术路线以及部分关键节点：

10-15

确定选择的原始数据集为成都二手房交易信息列表，对原始数据集的字段进行了分析和初步筛选

10-17

制定了初步的技术路线图

10-18

明确了要实现的目标为：买方、卖方预测系统；可视化页面

10-19

确定了使用网页的形式来完成我们的系统并进行展示

10-25

对数据进行了初步预处理，使用正则表达式、数值化等方式得到了经过第一次预处理的数据集

10-26

使用回归决策树的方式，基于十个特征对房价进行了预测，得到了合理的平均误差和方差。

10-27

开始使用经过第一次预处理的数据集进行可视化工作

11-02

进一步确定了可视化的具体展示形式

11-03

使用爬虫获取了时间数据

11-06

使用爬虫和百度地图API获取了地址和经纬度数据

11-11

发现爬取的地址存在一定格式上的问题，重新进行了处理，并找到和删除了部分异常数据，并进行了相关性分析，得到相关性的热力图。得到第三次预处理后的数据集

11-18

实现可视化页面点击中央大地图展示周围子图的功能。实现卖方、买方预测页面的底层功能。

11-19

实现网页后端

11-21

完成网页前端

11-22

对聚类进行了分析实践，对不同分类预测算法进行了评估，选择了效果最好的预测算法。

总结：

通过整个项目的实践，我们亲身体会了数据挖掘的那张路线图，预处理、分析之后发现问题（Knowledge），再进行新的处理，再重新分析挖掘，做评估，然后发现新的问题，再从头开始，在这几个过程的循环往复中完成了整个项目。

# 研发情况

## 数据预处理

1. 数据集描述

我们采用的数据集为采集自贝壳二手房交易平台的成都的二手房成交信息列表。数据集中一共有20001条数据，原始数据集每条数据有14个字段。

该原始数据集的14个字段中存在2个字段与内容无关/无效，存在2个字段需要深入获取信息，另外部分字段的内容需要拆解。

第一阶段我们选取的有效字段为['区','所处楼层级别','总楼层数','建造年份','面积','朝向','关注数','是否有VR房源','是否近地铁','单位价格','总价']

1. 数据提取和数值化

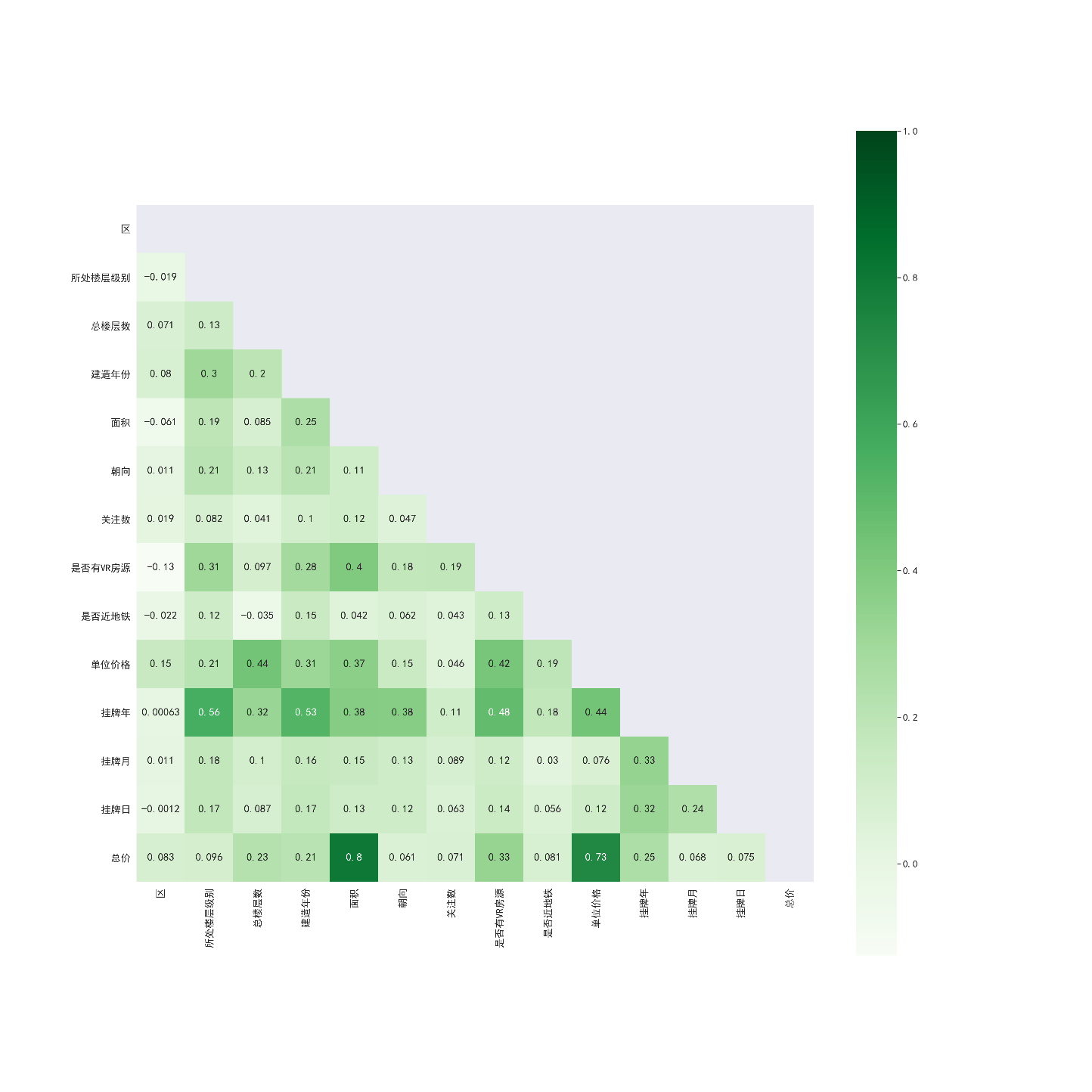
原数据集数据存在多个不同数据挤在一个字段中的问题，我们使用正则表达式对需要的字段进行提取，然后对必要的文本型数据进行数值化。

1. 爬虫

为了获取更多的数据，我们从原数据集的房屋链接字段爬取房屋的挂牌年月日；从小区链接字段爬取房屋所处小区的具体位置（文字描述）。然后申请百度地图的ak，通过百度地图将文字描述的位置转化为对应的经纬度。

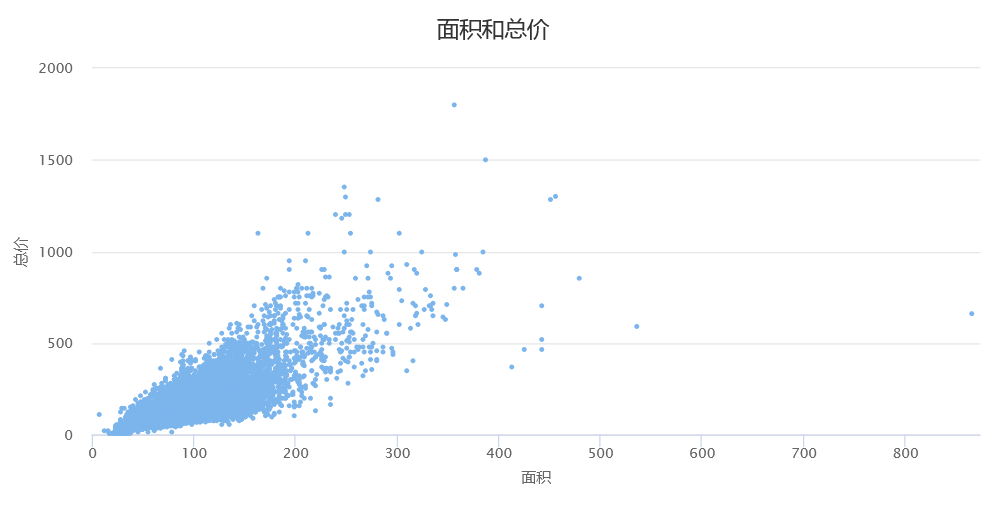
1. 相关性分析

通过相关性分析的方式，找出字段之前的关联性。



1. 异常值寻找

利用相关性分析，选取相关性高的两个特征（面积和总价），作出两个特征间的散点图：



可以看到在最右侧出现面积非常大而价格特别低的点，可以认为这是异常点删去；在最上方也出现面积较小但价格异常高的点，可以认为是异常点删去。

1. 数据处理结果

最终我们处理得到了两份数据，一份用于可视化，一份用于算法。具体数据说明如下：

processedDataForVisual的数据字段为：'编号','位置','区','房型','所处楼层级别','总楼层数','建造年份','面积','朝向','关注数','是否有VR房源','是否近地铁','单位价格','挂牌年','挂牌月','挂牌日','总价','纬度','经度'

processedDataForCalculate的数据字段为：'区','所处楼层级别','总楼层数','建造年份','面积','朝向','是否近地铁','挂牌年','挂牌月','挂牌日','纬度','经度','总价'

数值化数据字段说明

区：

{'双流': 0, '天府新区': 1, '成华': 2, '武侯': 3, '温江': 4, '郫都': 5, '金牛': 6, '锦江': 7, '青羊': 8, '高新': 9, '高新西': 10}

房型：

1|1指一室一厅，以此类推

所处楼层级别：

{'地下室':0, '低楼层':1, '中楼层':2, '高楼层':3}

朝向：

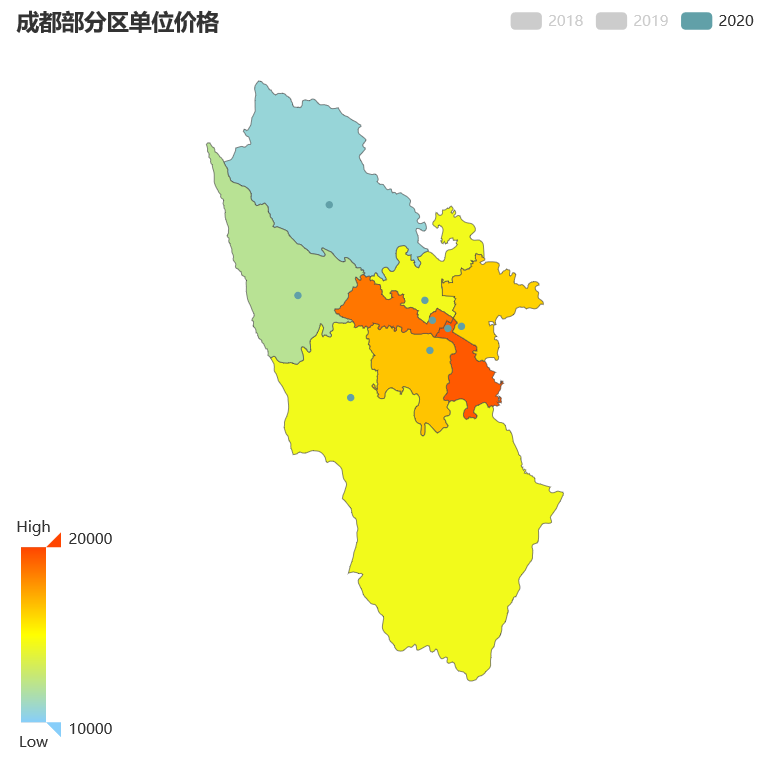
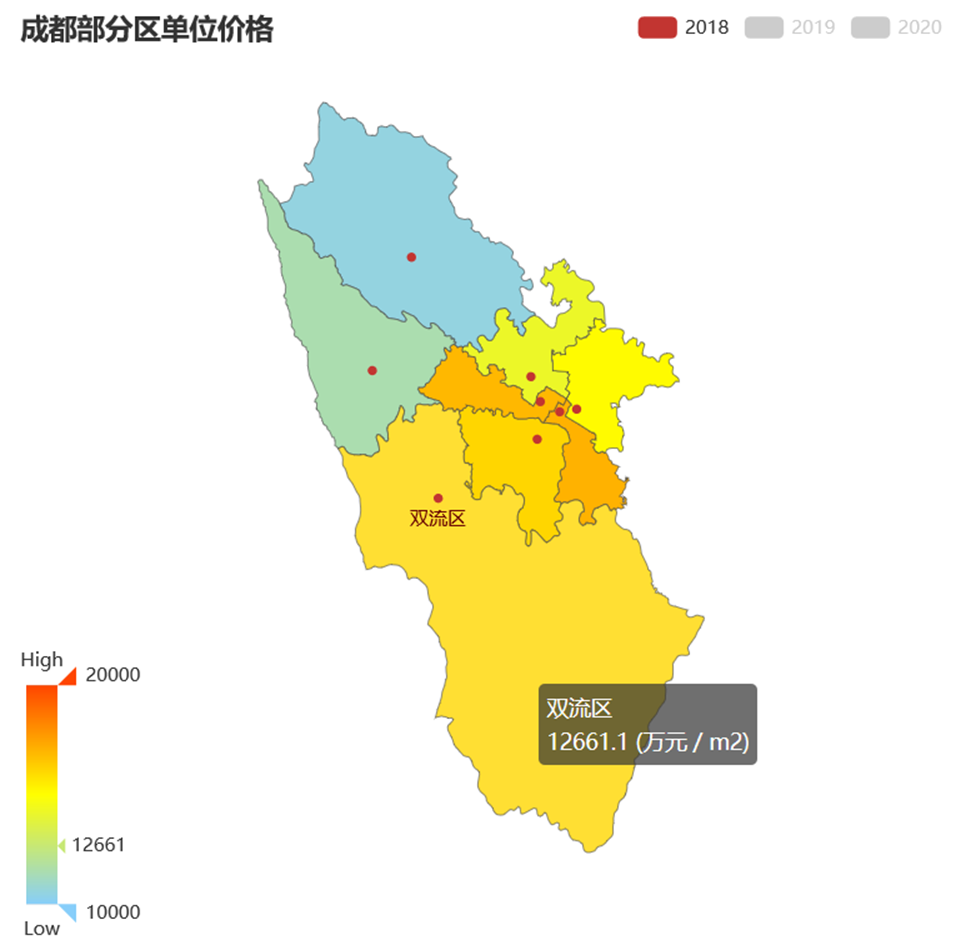
{'北':1, '东北':2, '东':3, '东南':4, '南':5, '西南':6, '西':7, '西北':8}

是否有VR房源/是否有地铁：

1表示是，0表示否

## 可视化分析

我们根据相关性以及感兴趣的部分做出大量图表，现对部分图表进行分析。



1. 中心城区（成华区、青羊区、锦江区、武侯区）单位房价更高。在正常的认知下，中心城区的房屋单位价格都要高于相对比较偏远的地区，在较大城市更加明显。我们以行政区为类别，计算量每个行政区房屋单位价格的中位数并标以颜色深浅。从图中可以看到，较靠近中心的几个行政区颜色都较外围行政区要高，在上右图（2020年数据）体现的尤为明显。
2. 单位房价在随时间升高。从2018年到2020年，肉眼可见的，每个行政区的颜色在加深，即单位价格在升高。每个行政区，即每个局部的单位价格在升高，可以合理地推断出由局部组成的总体的单位价格在升高。
3. 中心城区（成华区、青羊区、锦江区、武侯区）二手房的关注数更高，更多人关注这些地区的房屋。我们对每个区计算在该行政区内二手房的关注总数：

上方三张图为锦江、武侯、青羊三个中心行政区的总关注人数。

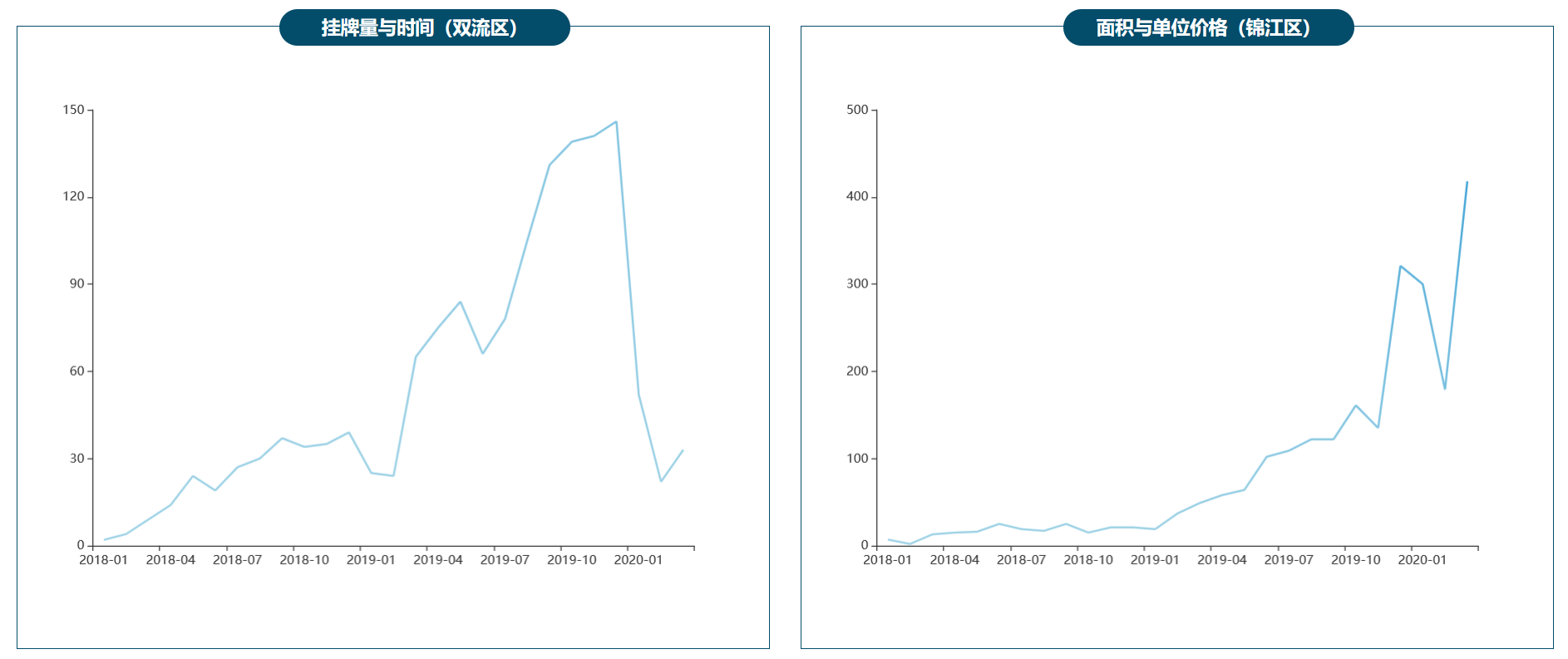
下方三个图为双流、郫都、温江三个较边缘行政区的总关注人数。

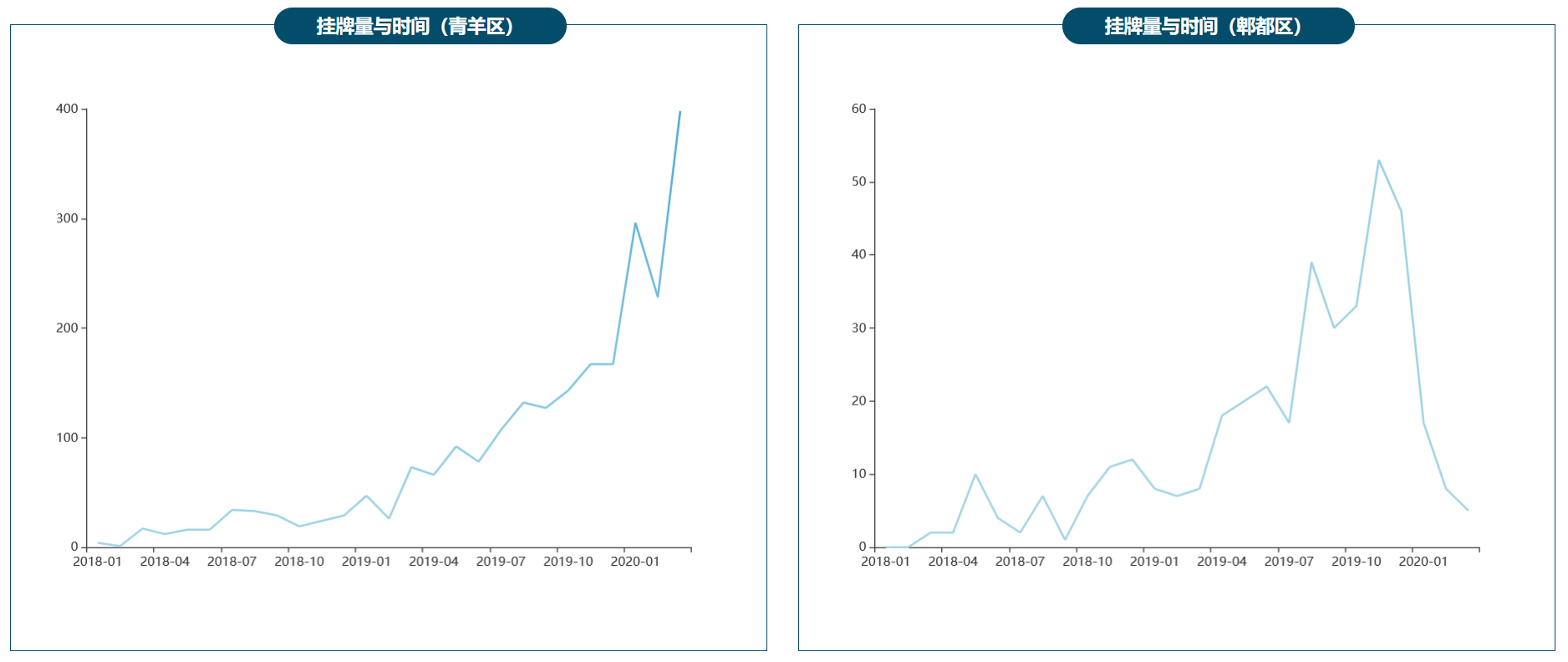
  

可以看到，中心行政区的总关注数要远远大于较边缘的区域。

我们认为这是由于中心区域在交通、地理位置、教育资源、娱乐等各方面都相对比较发达和完善所导致的。人们在选择房子时往往会考虑到周边环境。

1. 中心城区的二手房的挂牌量在进入2020年后飙升，而边缘城区的挂牌量则在进入2020年后显著下降。

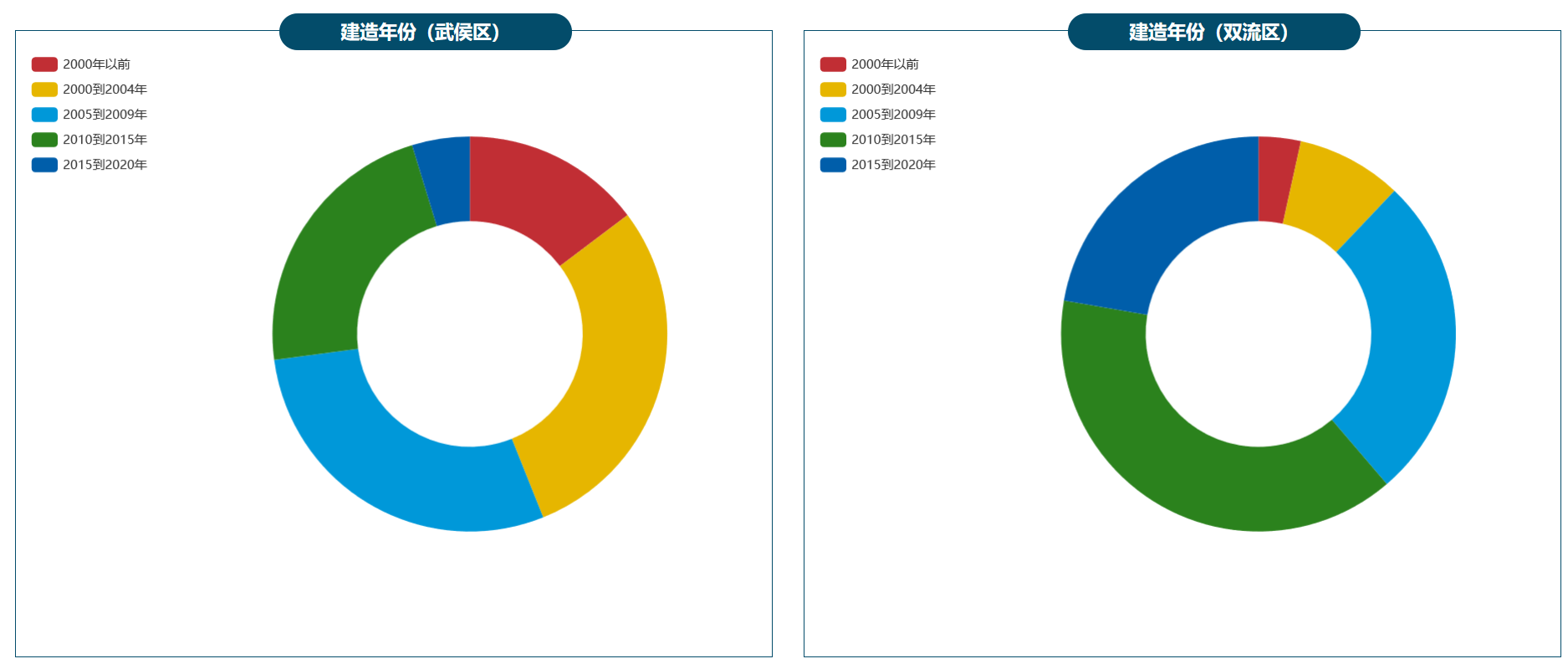




其中锦江和青羊均为地理位置上的中心区，郫都和双流为地理位置上的边缘区。

这一现象我们暂时无法解释，可能与政府政策有很大的关系，也可能是数据集本身的原因。但是我们依然认为这一现象有讨论的价值。

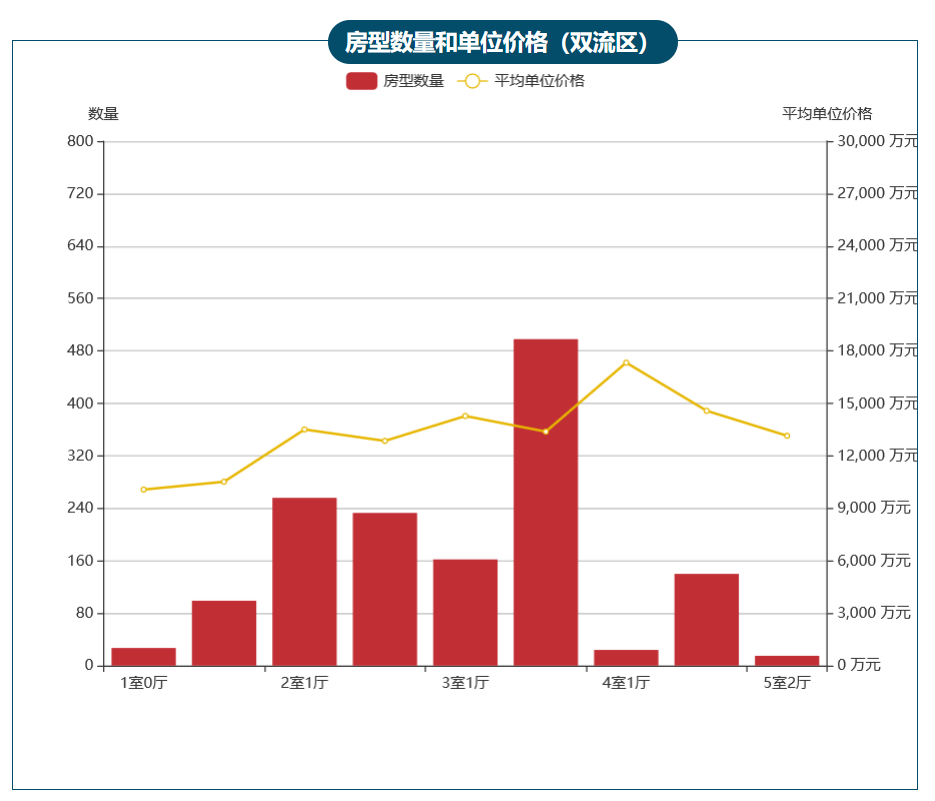
1. 中心城区的房屋建造年份普遍较早。

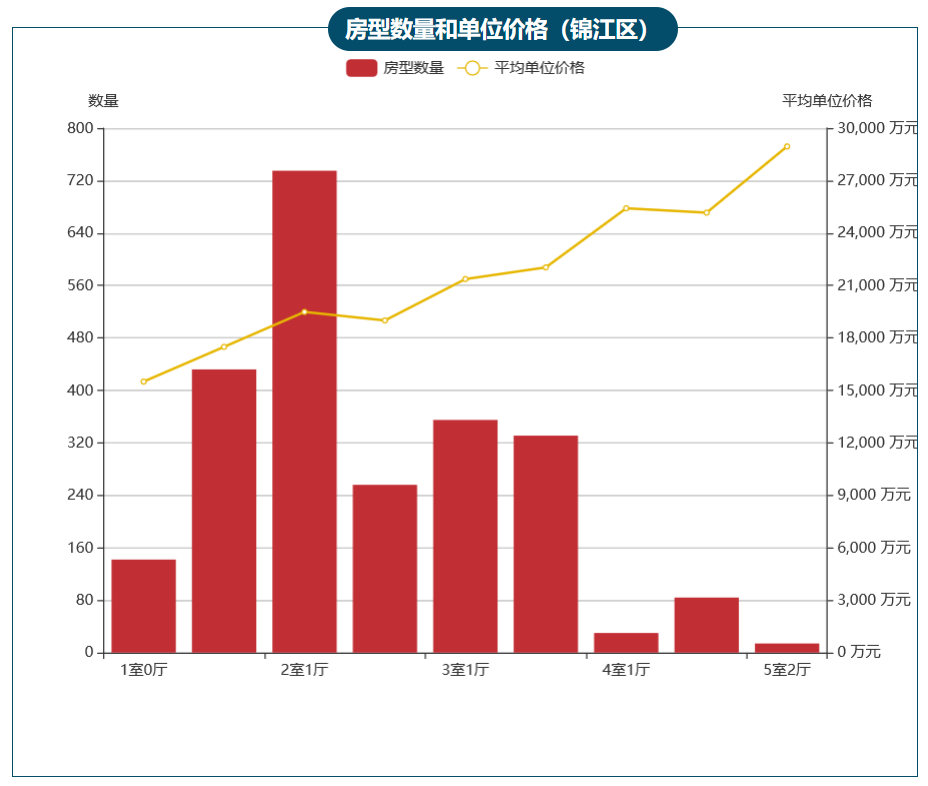


左图为武侯区。我们可以看到，武侯区在2005年以前建造的房屋比例（红色和黄色）要远远大于双流区，而双流区在2015年以后建造的房屋的比例（深蓝色）要远远大于武侯区。

这与成都市的历史有比较大的关系，锦江区、青羊区、武侯区大致在1953年就已经存在（当时为东、西城区），而双流等周边行政区是在1976、2016、2017才划分出来的。

1. 房型越大，单位价格越高：

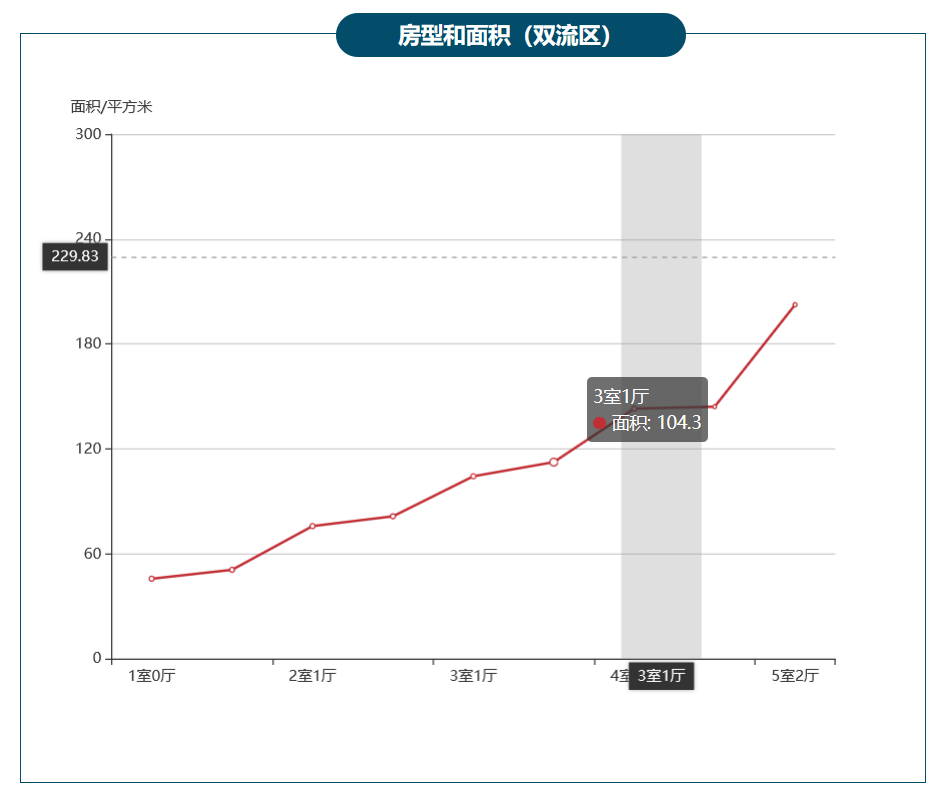




总体而言，随着房型的增大，厅室的增多，单位价格在逐步上升。在部分区的趋势很明显，而在少数几个区（例如双流）趋势比较不明显。

其原因大致可以这样推导：

首先，房型增大，面积在变大，这是符合实际情况和逻辑的：



其次，面积与单位价格的相关性为0.73，正相关，且相关性比较高，可以认为随着面积的增大，单位价格在增大。根据传递性，房型增大->面积增大，面积增大->单位价格升高，则房型增大->单位价格升高。

## 价格预测及聚类分析

1. 回归决策树

输入：'区','所处楼层级别','总楼层数','建造年份','面积','朝向','是否近地铁','挂牌年','挂牌月','挂牌日','纬度','经度’

输出：总价

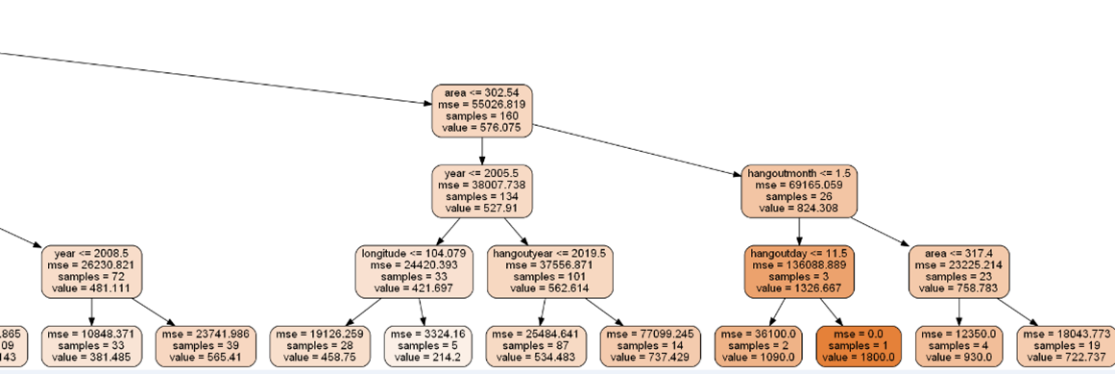
层数12

平均绝对误差 MAE = 35.54

均方根误差 RMSE = 57.83

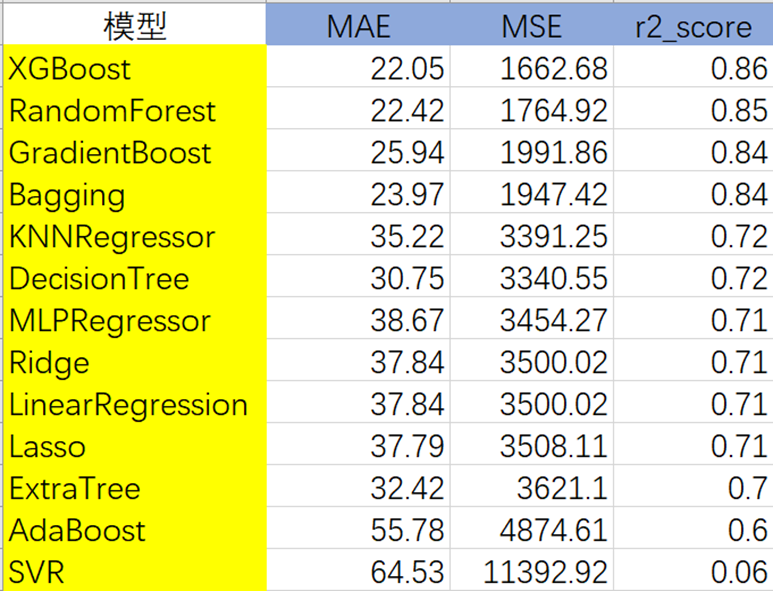
决定系数 r2\_score=0.72

决策树局部展示：（字段为英文）



1. 最优算法选择

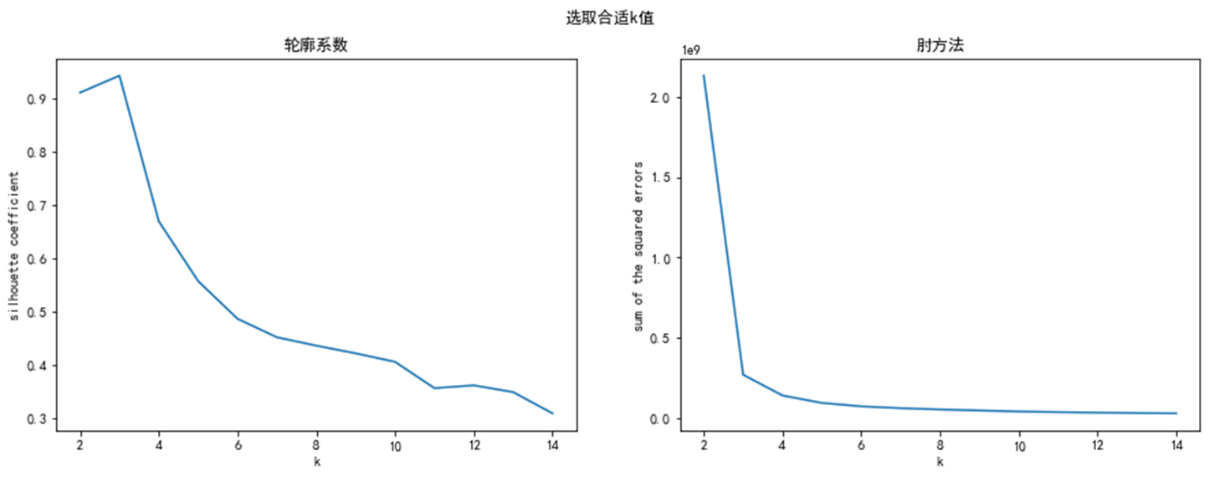
回归决策树的结果不尽如人意，为了进行更准确的预测，我们使用线性回归，KNN，支持向量回归，岭回归Lasso回归，神经网络，决策回归树，极端随机树，xgboost，随机森林, Adaboost,gradientboost，袋装等算法进行回归预测，使用平均绝对误差、均方误差、决定系数作为评价标准，以期选出最优算法。对算法按照评价标准从优到劣排序如下：



最优算法为xgboost

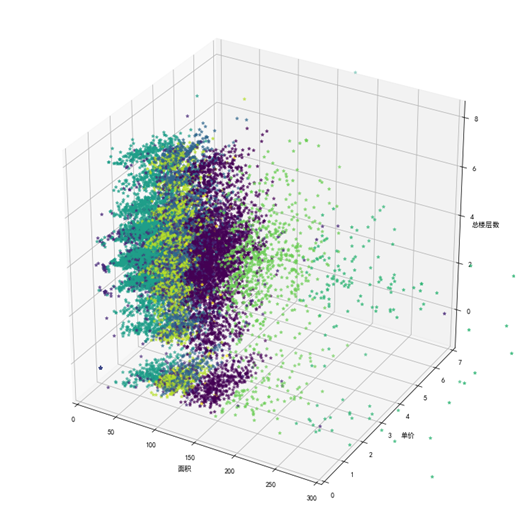
1. 聚类分析

首先找出合适的k值。使用轮廓系数和肘方法，作图如下：

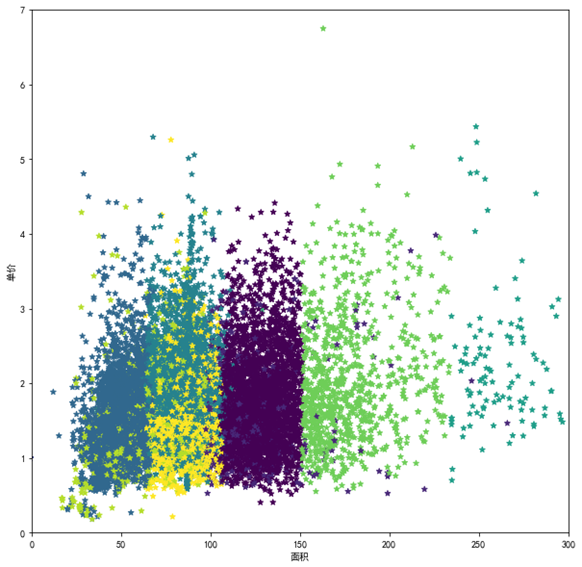


设置k=6，进行聚类。对聚类结果进行三维展示。

设置x轴为面积, y轴为单价，z轴为‘区’,‘所处楼层级别’,‘总楼层数’,‘建造年份’,‘面积’,‘朝向‘等属性中的一个。



经过对比发现聚类结果仅与面积和单价相关，与其它属性（z轴）均无关。因此使用二维图展示聚类结果。



对数据进行观察后，可以对聚类结果做如下分析：

小面积高价格的房屋多为市区旧建小宅，大面积高价格多为市区新建大宅，极大面积为豪宅，大面积低价格的房屋多为郊区偏远大宅，小面积低价格多为地段不好的小宅，极低价格为老旧小区住宅。

## 可视化页面

可视化页面的设计思路是：以中心的成都市的部分行政区划的地图为核心，以行政区划为类别，展示不同行政区划的各个图表。

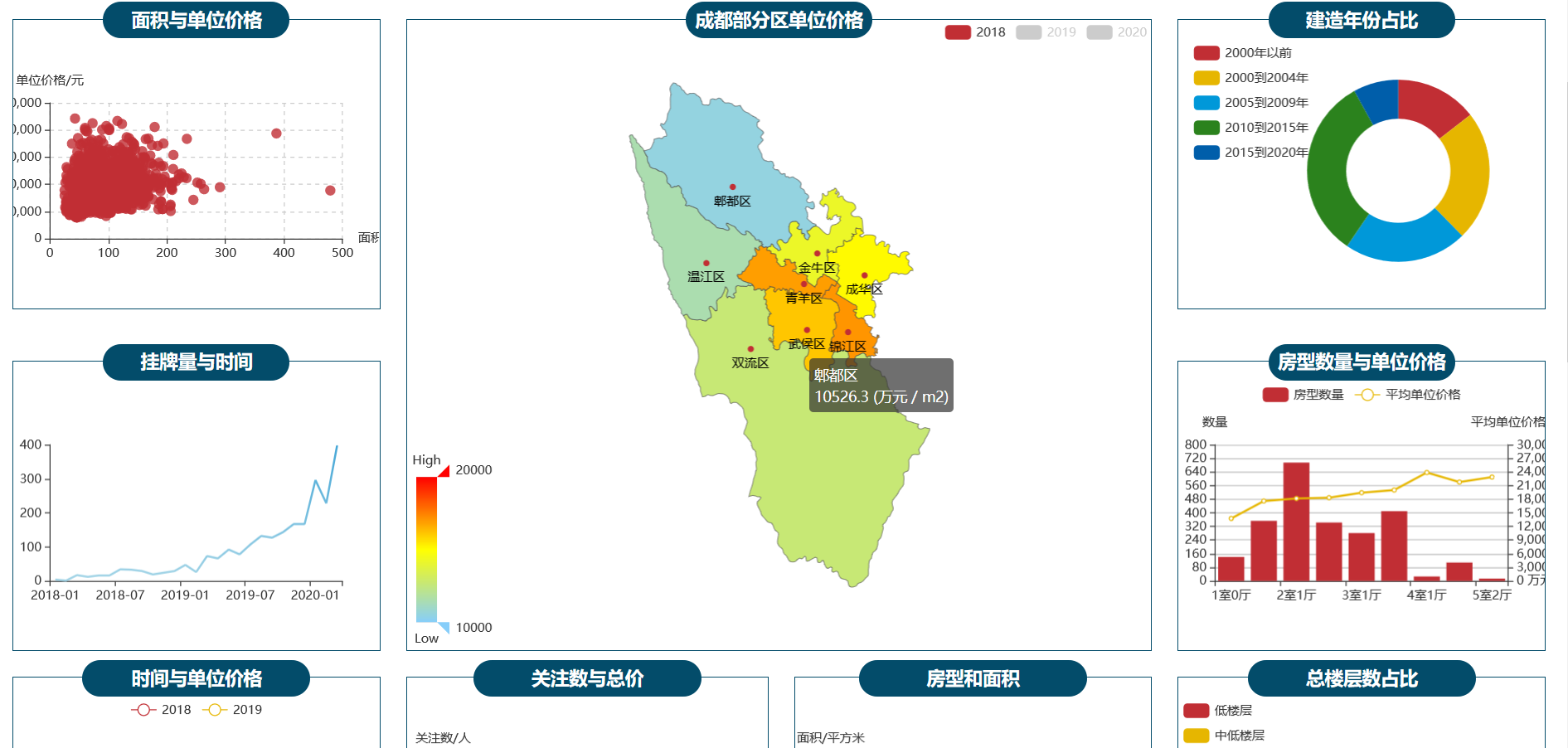
利用echarts的点击事件，使得用户可以通过点击某个行政区划，来讲该行政区划的图表展示在周围的子表当中。

同时，如果用户对某个子表产生兴趣，并想要做区之间的对比的话，可以点击子表中的任何一点，将想看的子表展示在下方的“大屏”上。

顶部信息栏：



中央区域：

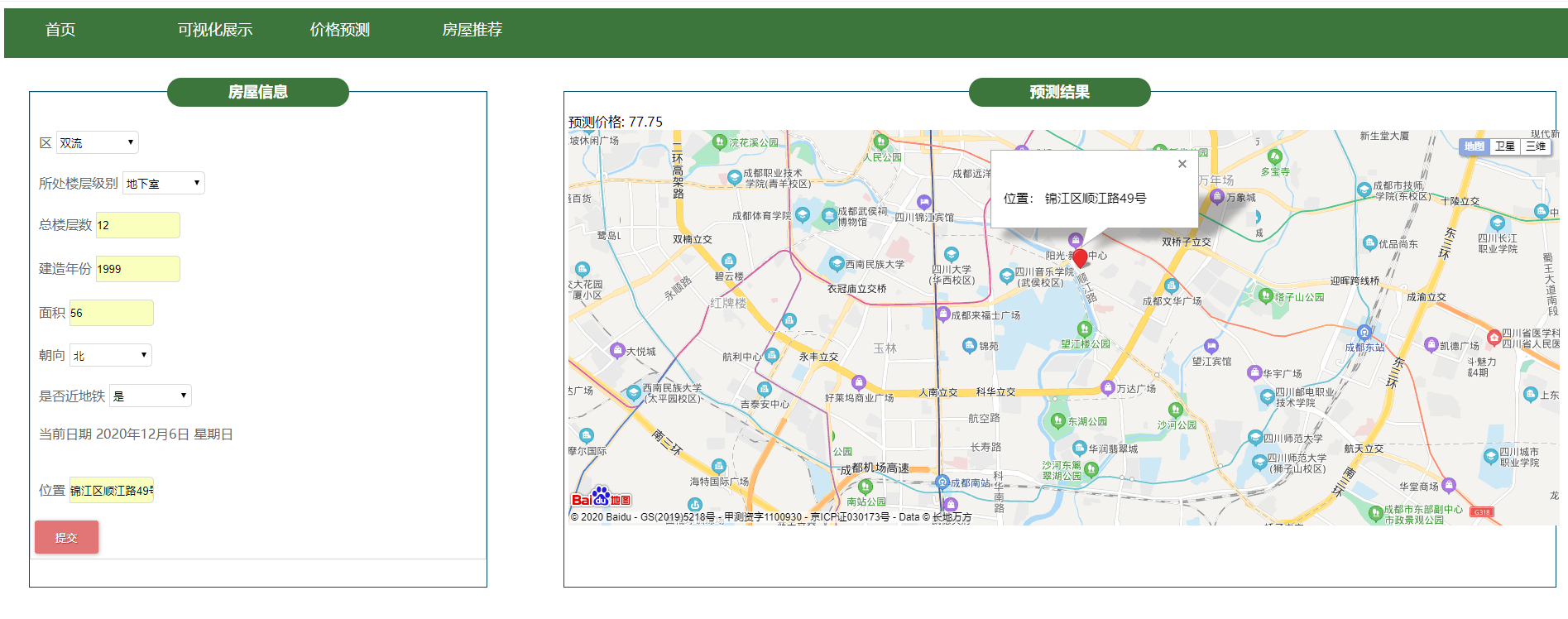
****

底部大屏展示区：



## 卖方价格预测页面

卖方在左侧表单内输入房子的各种属性（填空值可以缺省），点击提交，右侧显示预测价格，并在地图上将该房子标出来。



## 买方房屋推荐页面

买房在上方表单内输入自己有意向购买的房屋的属性范围（填空值可以缺省），点击提交，下方左边表格内显示系统推荐房屋的信息，右侧地图上标记出这些房子的位置。



# 成果总结

## 完成情况

以网站为载体，完成了以下功能：

卖方预测功能及买方推荐功能：

完成了基本的买方推荐功能，用户可以自主选取各个属性可接受的范围，系统会将符合用户需求的二手房在表格中展示出来并在地图上标注。

完成了卖方预测功能，用户可以提供需要售卖的二手房的属性，系统将会输出适合的价格并将需要售卖的房子在地图上标注。

数据可视化功能：

展示部分属性间的关系，包含了折线图、散点图、饼图等等。利用一张成都市部分行政区划的地图，通过点击地图可以展开更多的图表。同时可以将图表进行大屏和对比展示。

## 人员分工

徐尘化：数据预处理、数据可视化、网页前端

梅浩楠：数据预处理、爬虫、算法、网页后端(Django)

# 项目展望

希望能够实现不仅限于成都的二手房交易分析及预测系统，可以选择全国任意城市，后端实现对指定城市二手房交易数据的自动爬虫和预处理，前端相应的做出可视化展示，卖方价格预测功能和买方房屋推荐功能亦可实现复用