**《数据分析原理》**

**课程设计报告**

|  |  |  |
| --- | --- | --- |
| **组长** | 姓名 | 胡恒东 |
| 学号 | 5120192963 |
| **组员** | 姓名 | 陈瀚洋 |
| 学号 | 5120196330 |
| **组员** | 姓名 | 张传升 |
| 学号 | 5120195350 |
| **组员** | 姓名 | 刘璨竹 |
| 学号 | 5120196437 |

**指导老师: 骆蓉**

**2021年12月**

**杭州市二手房房价分析**

**摘要： 本次是对杭州市二手房价的分析，通过对相应的房价标签进行可视化图表统计并的出相应结论，同时使用回归模型对房价进行预测评估。**

**关键词：3-5个**

**杭州市二手房、房价指标分析、房价回归模型、二手房价分析**

目录

[一、数据分析目标与任务 1](#_Toc88496250)

[二、数据预处理 1](#_Toc88496251)

[三、数据探索分析 1](#_Toc88496252)

[四、数据分析模型 1](#_Toc88496253)

[五、方案评估 2](#_Toc88496254)

## **一、数据分析目标与任务**

1. 背景介绍

随着我国房地产业的迅速发展，城市可供开发的土地越来越少，二手房交易日趋活跃，蕴藏着广阔的发展空间与极大的发展潜力。由于二手房市场情况比较复杂，如何快速且科学得为二手房定价一直是众多学者研究的重要课题，它不仅能保障二手房交易双方有效地推进业务，实现交易双方的利益最大化，还能为二手房市场的长期健康发展奠定坚实的基础。

1. 课程设计目标与任务
2. 获取链家网杭州市的二手房数据
3. 对数据进行预处理
4. 分析影响该城市二手房价格的主要因素
5. 构建相应的二手房房价估计模型以及评估、优化模型
6. 研究方法与技术路线

**数据获取：**通过python的requests库进行网页内容获取，re库和lxml库的etree进行数据解析，以及

pandas库进行数据存储和time库实现间歇爬取从而实现反反爬。

**数据预处理：**主要通过pandas以及numpy库进行实现。

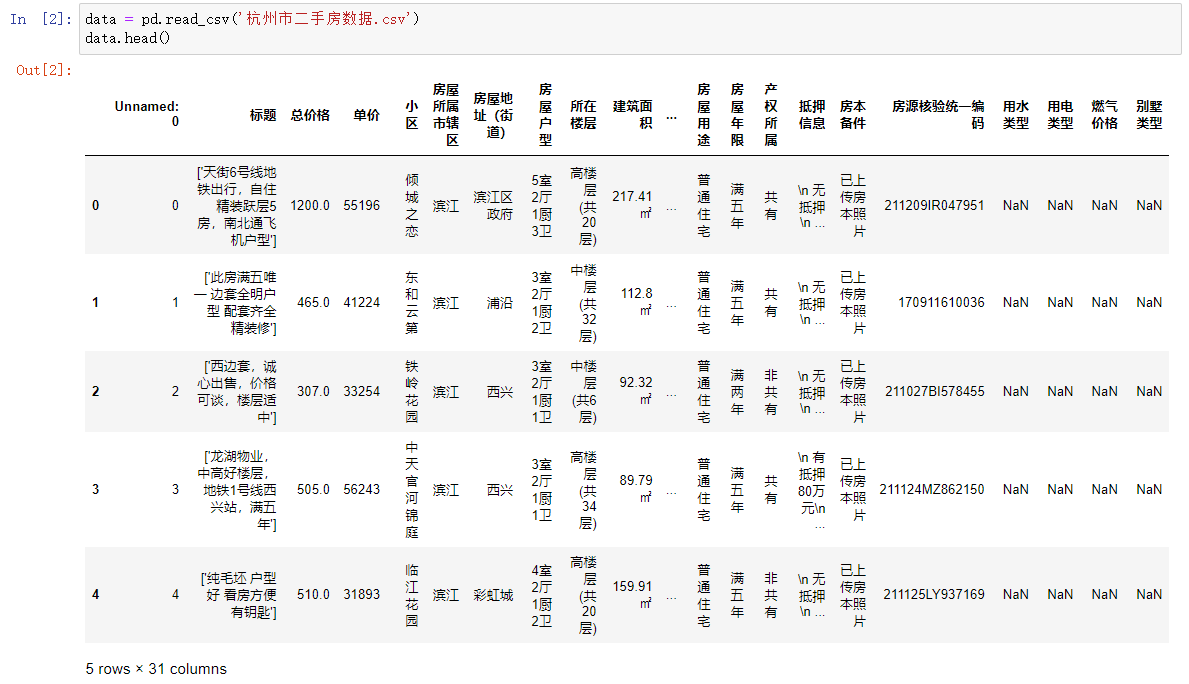
**数据分析：**主要通过matplotlib、seaborn库进行画图分析。

**构建、评估、优化模型：**主要通过sklearn库进行实现

## 二、数据预处理

1. 数据说明

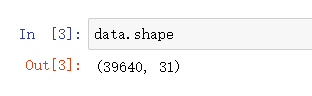
数据样本：



数据字段信息：

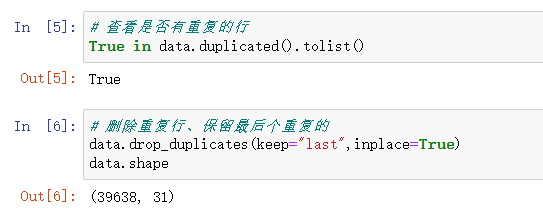
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **字段** | **说明** | **字段** | **说明** | **字段** | **说明** |
| 标题 | 售房标题 | 户型结构 | 如：跃层 | 交易权属 | 如：商品房 |
| 总价格 | 房屋总价 | 套内面积 | 套内房屋使用空间的面积 | 上次交易 | 如：2008/10/5 |
| 单价 | 每平方米单价 | 建筑类型 | 如：板楼 | 房屋用途 | 如：普通住宅 |
| 小区 | 住房所属小区 | 房屋朝向 | 如：南北 | 房屋年限 | 如：满五年 |
| 房屋所属市辖区 | 房屋所属市辖区 | 建筑结构 | 如：钢混结构 | 产权所属 | 如：共有 |
| 房屋地址（街道） | 房屋地址（街道） | 装修情况 | 如：精装 | 抵押信息 | 如：无抵押 |
| 房屋户型 | 如：5室2厅1厨3卫 | 梯户比例 | 如：两梯两户 | 房本备件 | 如：已上传房本照片 |
| 所在楼层 | 如：高楼层 (共20层) | 配备电梯 | 如：有 | 房源核验统一编码 | 如：211209IR047951 |

数据规模：

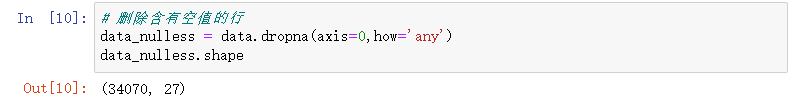
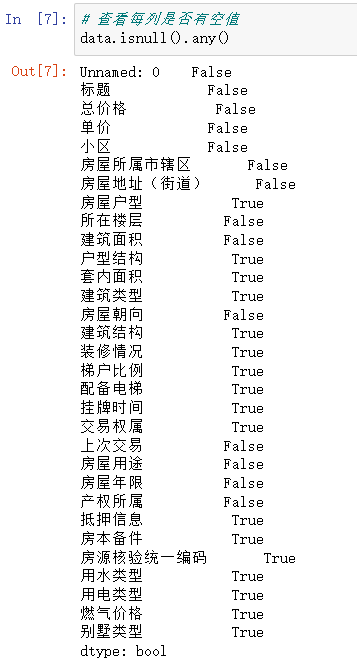


1. 数据清洗

重复行处理：



缺失值处理：



异常值处理：

对于房源核验统一编码列中，存在http://...，这样的连接数据，但是房源核验统一编码这一列对于数据分析并不太重要，这儿考虑将这列删除。

以上是数据清洗的主要步骤及原理，效果理想，保留下了唯一且有效的数据。

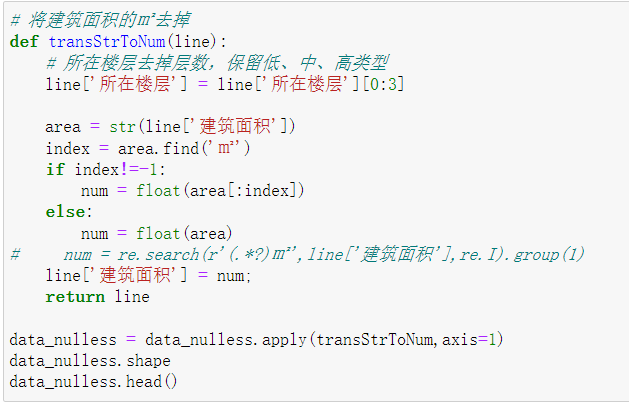
1. 数据处理

在数据预处理阶段，主要进行了数据类型的转换。见下图。

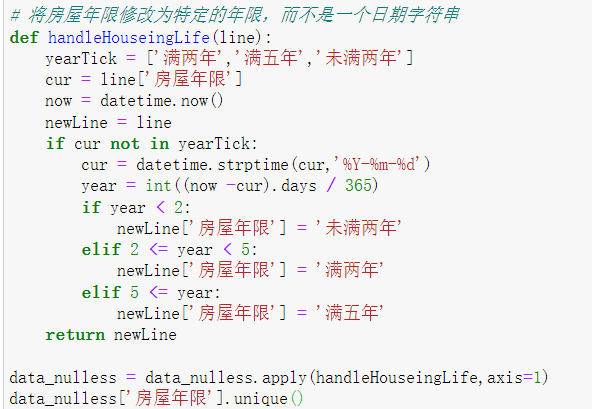
其他的相应的数据处理针对下面数据分析时针对具体分析做的相应数据处理。

基本上没有太大难点，主要就是对数据不太熟悉，需要在后面数据探索分析时逐步熟悉及解决。

去掉建筑面积列的单位，将其变换为float型：



修改房屋年限为日期字符串的，将其转为特定的类别：

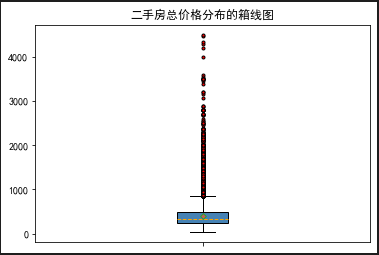


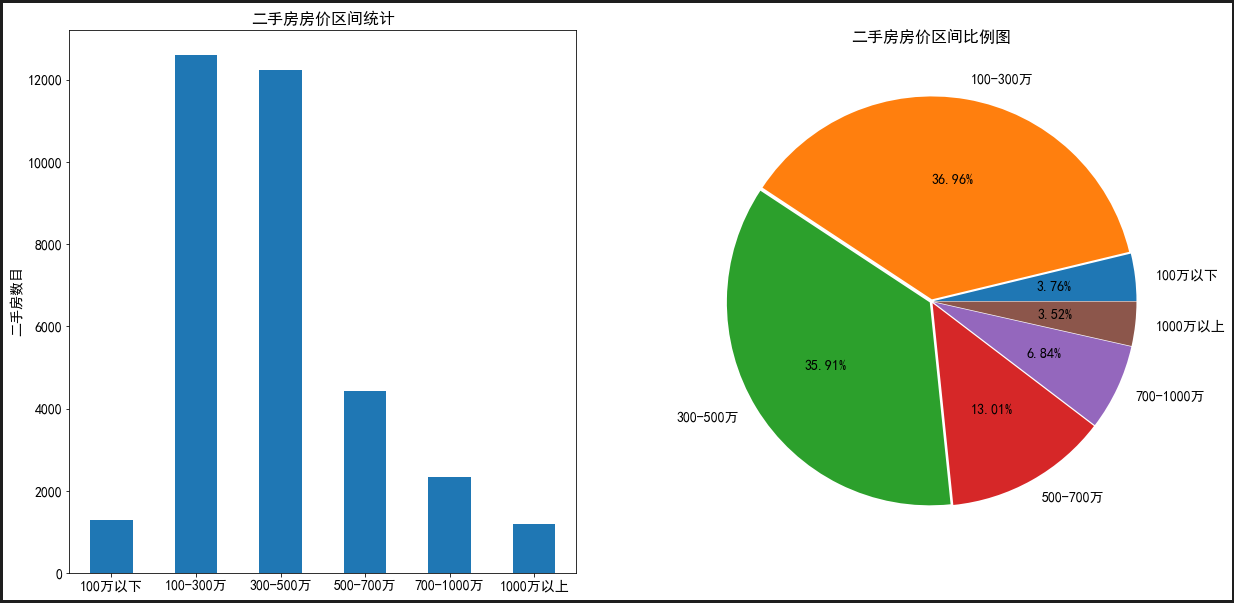
## 三、数据探索分析

1. 结合可视化呈现，对数据进行探索性分析

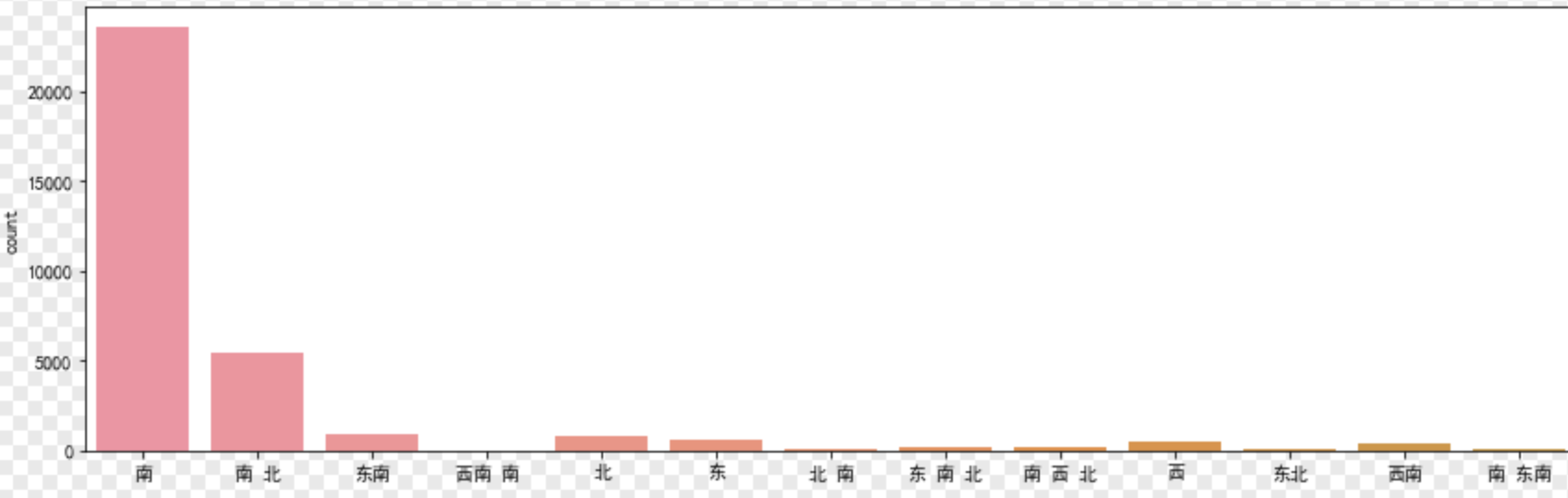
我们小组在观察数据集结构后，决定分别从建筑结构、房屋朝向、装修情况、所在楼层、房屋用途、房屋所属市辖区、房屋户型等特征值入手，分析这些特征值与二手房房屋价格的关系。

1. 可视化呈现结果



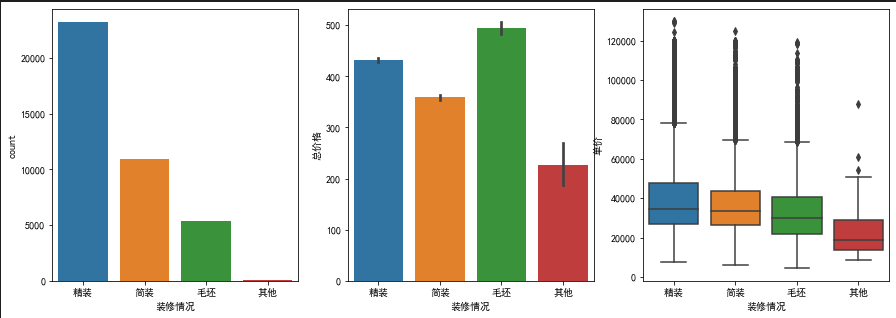


首先观察二手房总房价的大致情况，发现36.96%二手房的总价格在100-300万，35.91%二手房的总价格在300-500万之间，也就是100-500万的二手房价格是最普遍的，100万以下和1000万以上的二手房数量较少，分别占3%左右。

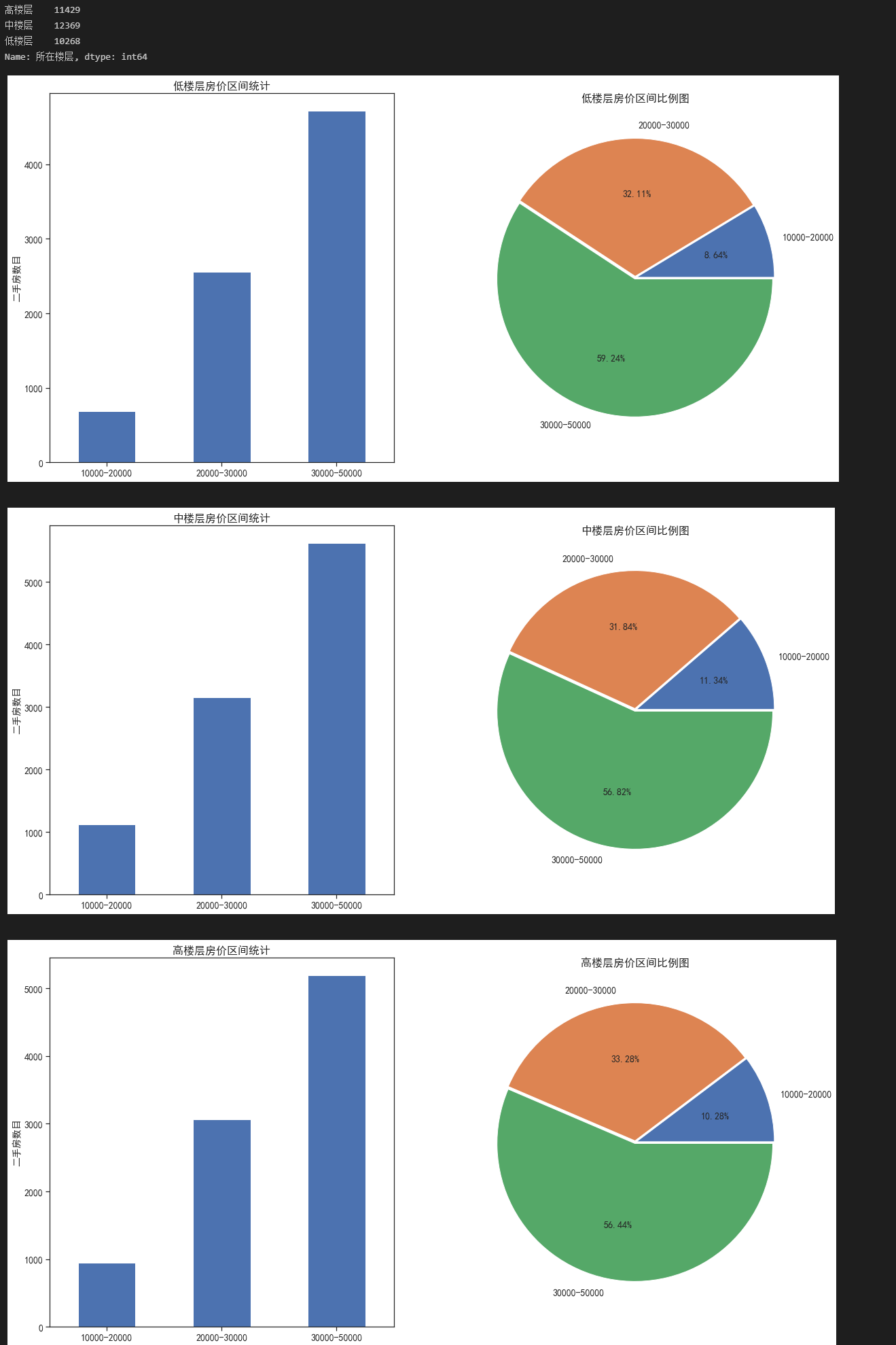




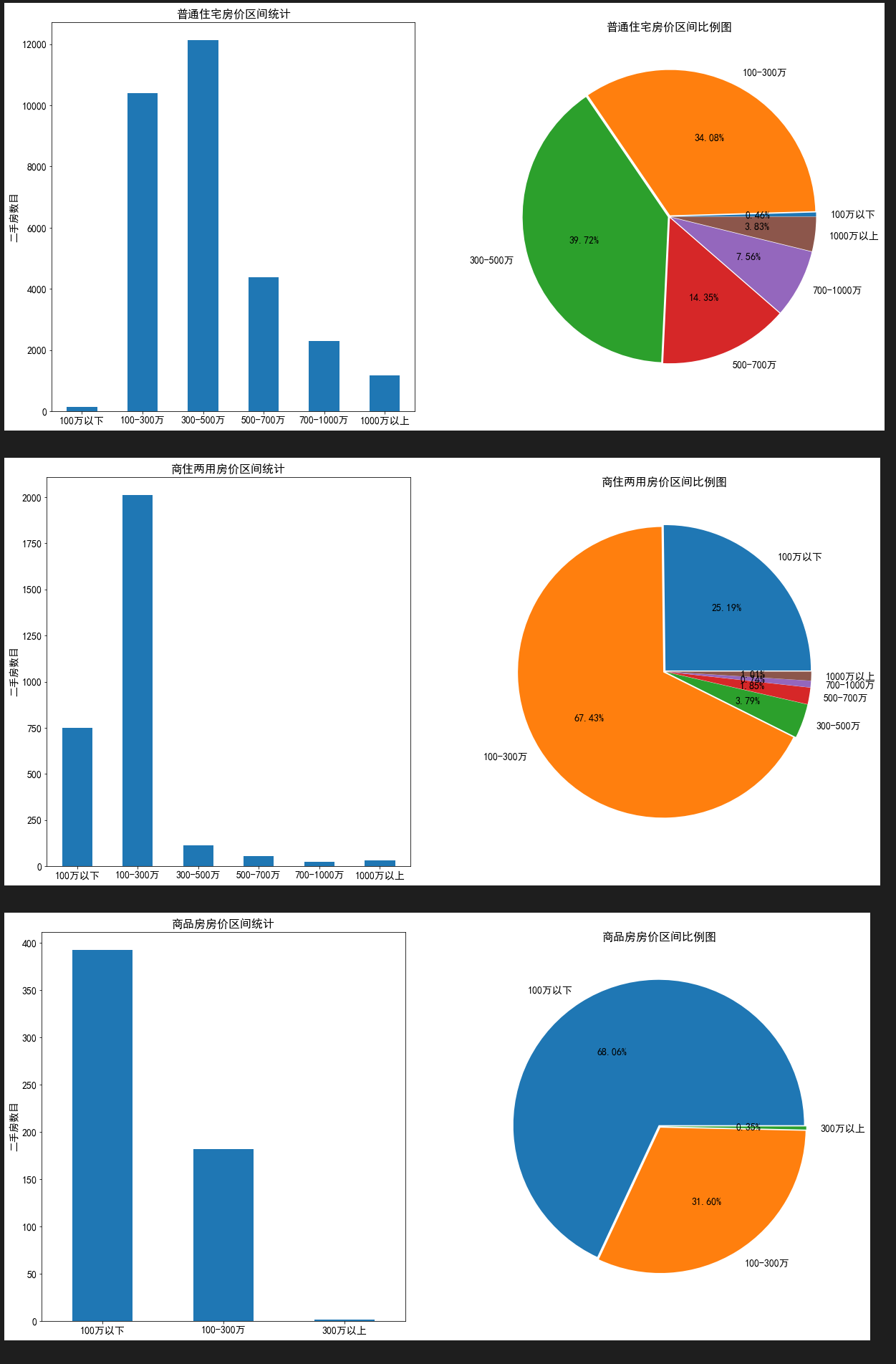
大多数二手房的房屋朝向南方，采用钢混结构、混合结构、砖混结构以及框架结构的二手房数目居多，采用钢混结构的二手房数目是最多的。



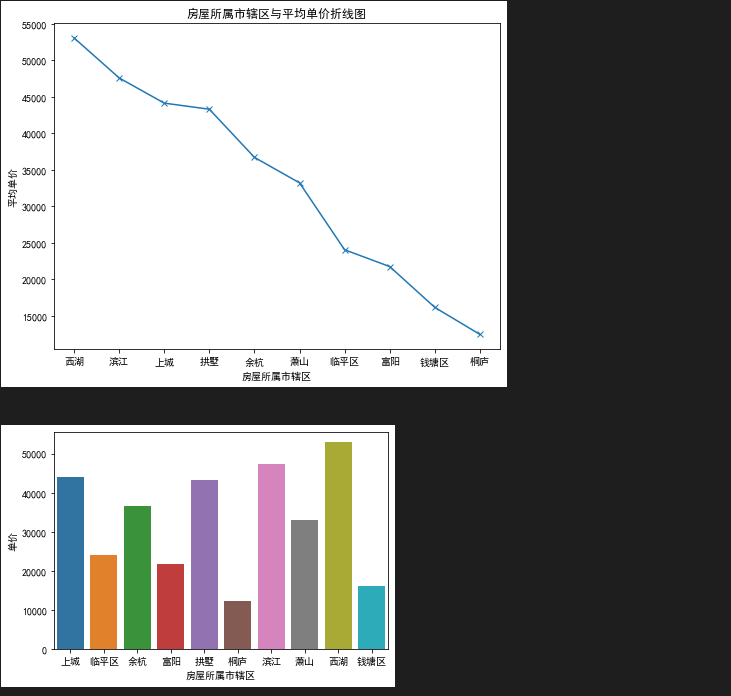
装修情况也是影响房屋价格的一个重要因素，我们大概可以先猜想一下，在日常生活中，装修较好的房屋一般是属于较贵的，而没装修的房屋呢？一般会便宜很多。那么从数据中，我们可以看出，二手房中，大多数是属于精装的，而毛胚房的数目是最少的。第三个箱型图中，就验证了该猜想，而第二个柱状图显示的总价格与装修情况的关系和我们预想的有很多差别，到底是为什么呢？因为总价格是面积乘以单价，那么由于第三个箱型图也验证了我们的猜想，就只能说明，毛坯房的面积是普遍较大的，所以比精装房普遍贵一点。



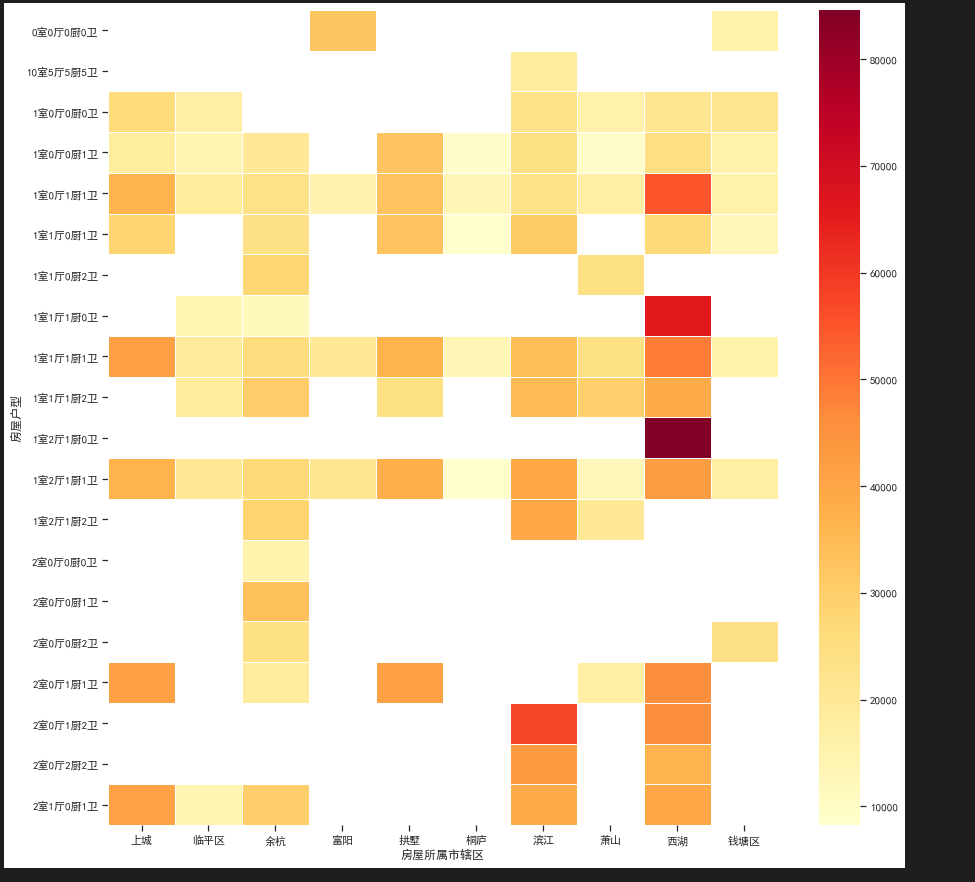
这里是所在楼层对二手房单价的影响。从图上大致可以看出，房屋的所在楼层对房屋单价的影响不大，不论是低楼层，中楼层，高楼层的房屋，他们的房屋单价分布基本一致，并无大的改变。30000-50000元/平方米的二手房居多。中楼层的房屋单价低价要比其他两个楼层多一点，10000-20000元/平方米的比例为11.34%，高楼层和低楼层分别为10.28%、8.64%。低楼层的房屋30000-50000元/平方米所占比例较高59.24%，所以中楼层的房屋可能会稍微便宜一点点，低楼层的房屋会稍微贵一点。



首先我们先查看了房屋用途类型，发现有商品房、商住两用房、普通住宅、平房等四种类型，但是由于平房只有一条数据，所以我们决定放弃绘图，而是直接打印出来查看的方式解决。而其他类型用途的二手房情况，如上图所示。最贵的还是属于普通住宅，总价格基本在100-500万之间。商品房是最便宜的，大多数二手商品房处于100万以下。然后就是商住两用房，大多处于100-300万这个价格段。



然后我们对房屋所属市辖区的单价进行了分析。最贵的房屋处于西湖区，平均在50000元/平方米，其次是滨江、拱墅、上城等40000元/平方米的价格。最便宜的属于钱塘、桐庐等10000-30000元/平方米。



由于数据量庞大，这里只选取了一部分数据构建热力图。其实这里和上一个可视化呈现大致是一样的，只是加入了户型这一特征值。在杭州这一列上，颜色基本很深，表示了杭州的房价很贵。越靠近下方的列其颜色一般更深，说明户型越大，基本房价就越贵。最贵的是西湖区的1室2厅1厨0卫的二手房，单价高于70000元/平方米。

1. 结论

完成数据分析之后，我们来做一个总结。从分析结果中可以看出，将普通住宅、精装，大面积和较发达的市辖区（如西湖）等词语组合在一起就构成了高价格的二手房，处于50000元/平方米以上的单价范围。将商品房、毛胚房，小面积和不怎么发达的市辖区等词语组合在一起就构成了略微低一点价格的房，但由于杭州本身作为一线城市，其相对低价格的二手房其价格也是处于10000-30000元/平方米的范围内，依旧处于高价格范围。

## 四、**数据分析模型**

1. 结合分析的目标，拟采用哪一种模型（如聚类、分类、回归）开展分析

这儿拟采用的是回归模型。对于房价预测案例来说，回归模型比较合适，其介绍为回归模型是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。这种技术通常用于预测分析，时间序列模型以及发现变量之间的因果关系。

其主要的方法有如下6种：

1. 随机森林回归
2. 支持向量回归
3. 梯度提升回归
4. KNR最邻近回归算法
5. 回归决策树
6. 线性回归

模型的标签选取：通过上述的探究分析，回归模型选取的标签主要有如下标签：

配备电梯，房屋所属市辖区,装修情况,房屋年限，总价格,单价,建筑面积。

对于字符型标签数据，进行哑变量处理：



1. 模型评估

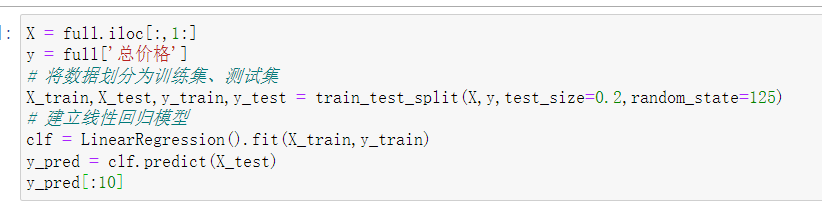
这儿主要用如下的回归模型评估方法：

平均绝对误差(mean\_absolute\_error),均方误差(mean\_squared\_error),

中值绝对误差(median\_absolute\_error),可解释方差值(explained\_variance\_score),

R平方值(r2\_score)

建立线性回归模型：



针对上面的线性回归模型进行指标评估得出：

线性回归模型的平均绝对误差为： 46.54674435910352

线性回归模型的均方误差为： 6533.38277479152

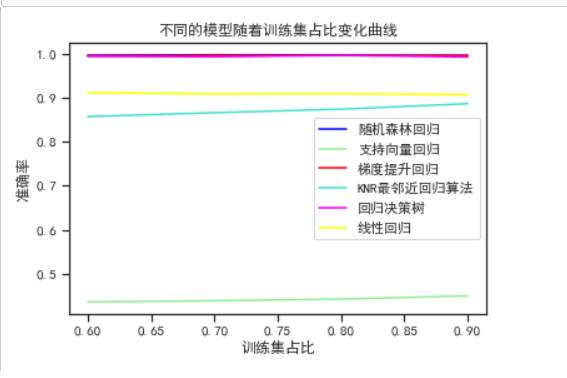
线性回归模型的中值绝对误差为： 28.13013693317629

线性回归模型的可解释方差值为： 0.9117924411939502

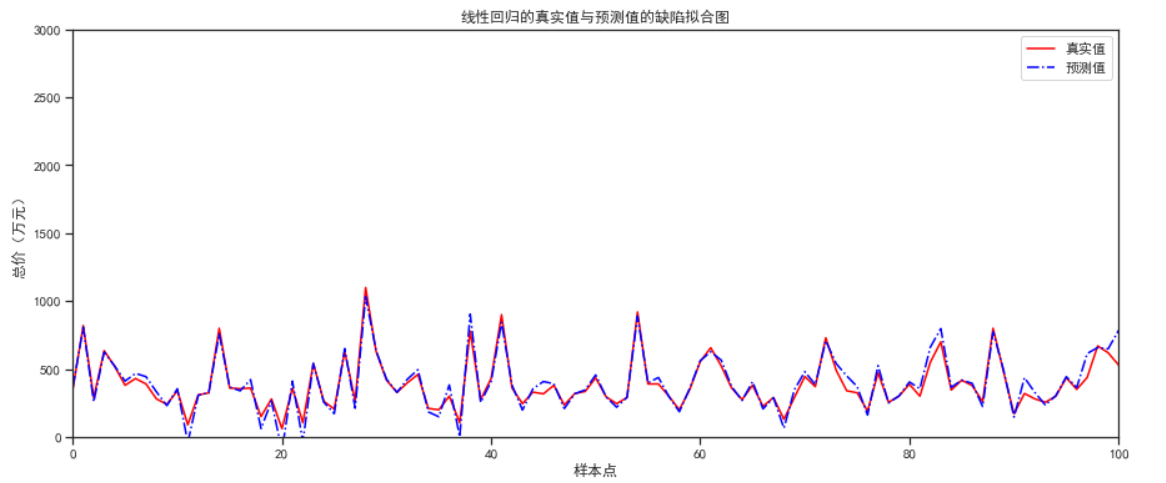
线性回归模型的值为： 0.9117316102038063

1. 可视化结果

下面是不同回归模型的训练集占比与准确率的关系折线图，从图中可见，几种回归模型对于训练集占比的影响不是很大，而对于回归算法来说，除了支持向量回归外，其余几种的预测结果的准确率都比较高。



从下图的真实值与预测值的拟合曲线来看，可以看出回归结果比较良好，有少数的点有一点偏差。



## 五、方案评估

本次课程设计的预期目标以及设计任务都已完成，课程设计的难度中等偏上，如果说还有要需要改进的话那就是可以进行更加深入的数据探索，从更大的维度上去观察分析。

## 附录

表 小组成员分工表

|  |  |  |  |
| --- | --- | --- | --- |
| **成员1** | **姓名** | **学号** | **专业班级** |
| 胡恒东 | 5120192963 | 卓计1901 |
| **主要完成工作** | 数据处理，数据分析图表绘制，房价回归模型建立与评估。 | | |
| **成员2** | **姓名** | **学号** | **专业班级** |
| 陈瀚洋 | 5120196330 | 卓计1901 |
| **主要完成工作** | 数据处理，数据分析图表绘制。 | | |
| **成员3** | **姓名** | **学号** | **专业班级** |
| 张传升 | 5120195350 | 卓计1901 |
| **主要完成工作** | 数据处理，数据分析图表绘制。 | | |
| **成员4** | **姓名** | **学号** | **专业班级** |
| 刘璨竹 | 5120196437 | 计科1901 |
| **主要完成工作** | 数据处理，数据分析图表绘制。 | | |