厦门二手房价预测以及分析报告

1. **研究背景**
2. **研究背景**

根据国家统计局发布了2019年12月份70个大中城市商品住宅销售价格变动情况统计数据。根据数据显示，12月厦门新建商品住宅价格同比上涨3.9%，环比上涨0.6%。涨幅位列全国第22，二手房价格环比上涨0.8%，同比上涨6.1%。涨幅居全国第8位，但是厦门并不是表面上看起来这么火热，厦门楼市呈现出了冷热不均的情况，岛内火爆，岛外部分区域遇冷的特点。在此背景下，本报告着手于分析厦门整体的房价分布，以及通过机器学习的方法预测厦门的房价并分析影响房价主要的因素。

1. **研究问题**

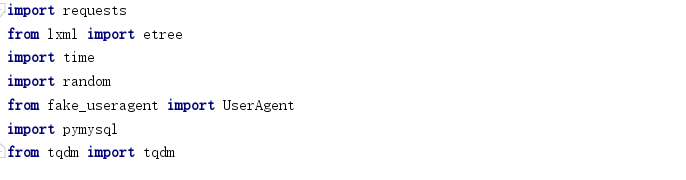
本报告主要的研究问题是研究分析影响厦门房价的主要因素。

1. **拟采用的技术**
   1. 网络爬虫
   2. Mysql数据库存储
   3. 机器学习算法
2. **数据采集以及存储**
3. **数据来源**

本次研究的数据主要来源于链家官网厦门二手房数据，数量为21000条，数据内容包括：房价、地址、小区名称、年份、楼层等特征

链家官网：[**https://xm.lianjia.com/ershoufang/**](https://xm.lianjia.com/ershoufang/)

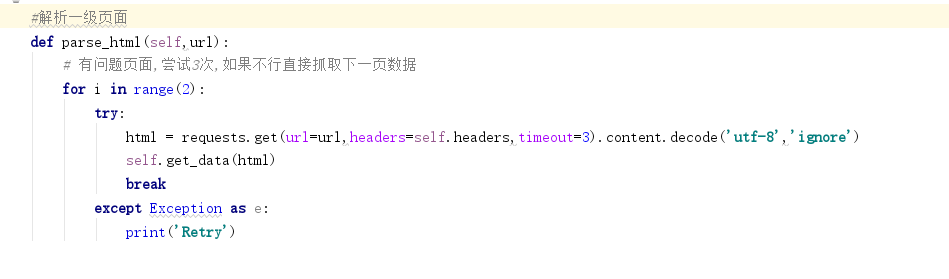
1. **网络爬虫代码**
   1. 导入所需使用的库



* 1. 创建连接地址以及连接数据库



* 1. 解析一级页面



* 1. 抓取一级页面数据



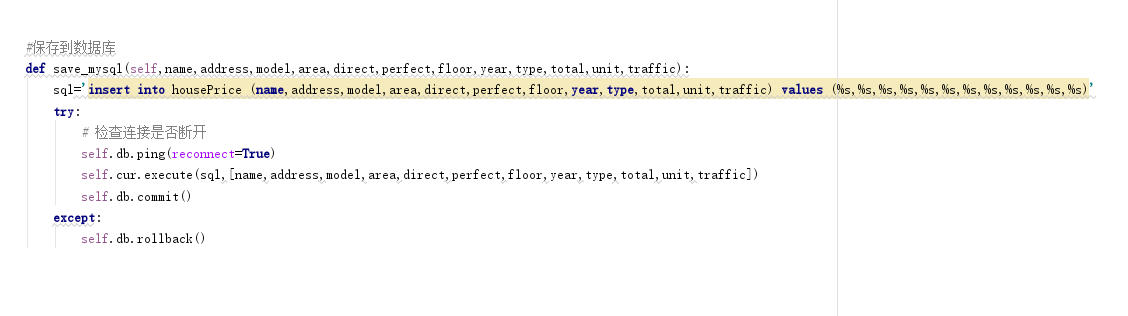
* 1. 解析二级页面



* 1. 获取页面数



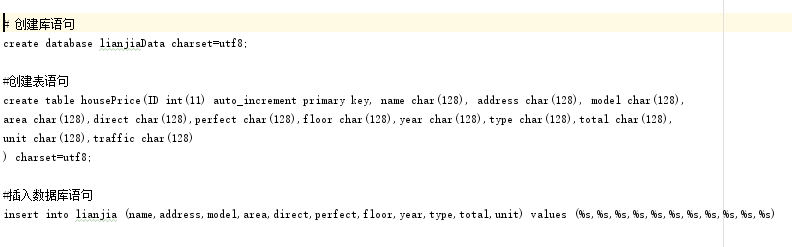
* 1. 保存到数据库



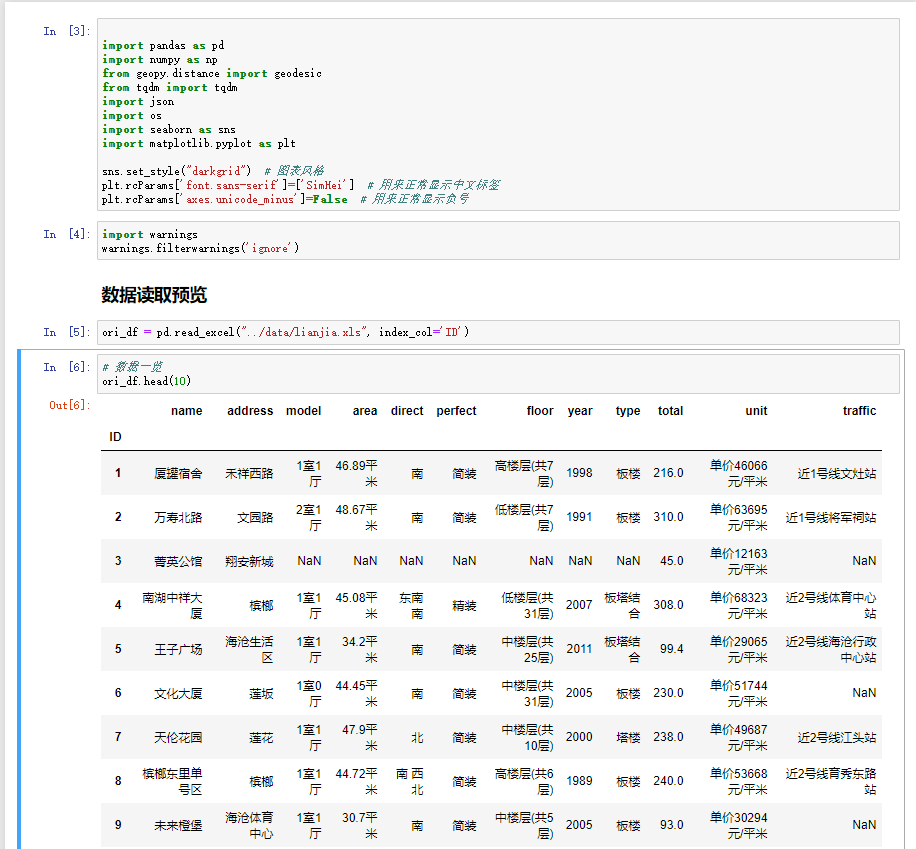
* 1. 运行主程序，启动爬虫，因为链家限制只能爬取3000条数据，所以采取了通过筛选条件分区域的方法爬取，避开反扒。



1. **数据库以及数据存储**
   1. 数据库采用的较为通用的mysql数据库，此数据库为关系型数据库。
   2. python通过pymysql模块，储存数据到数据库
   3. 数据库设计以及写入语句



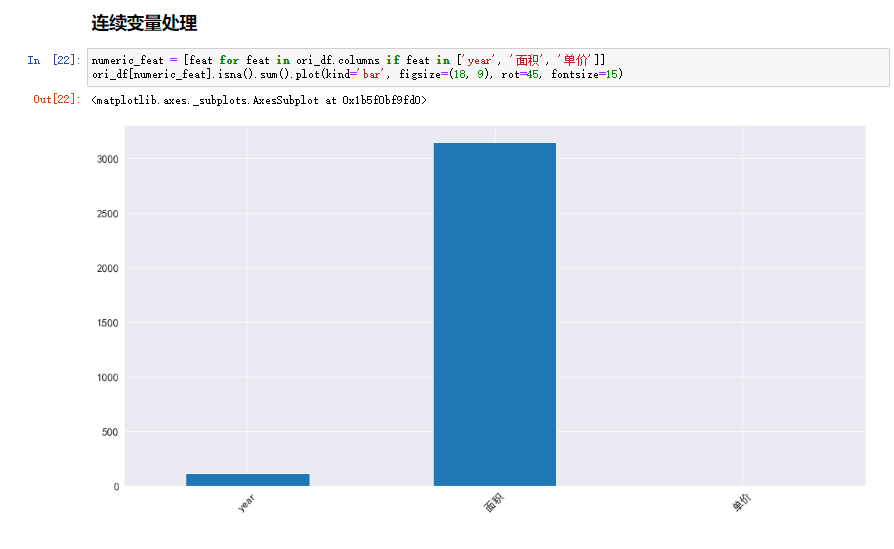
1. **数据集介绍、数据预处理、数据可视化、数据挖掘**

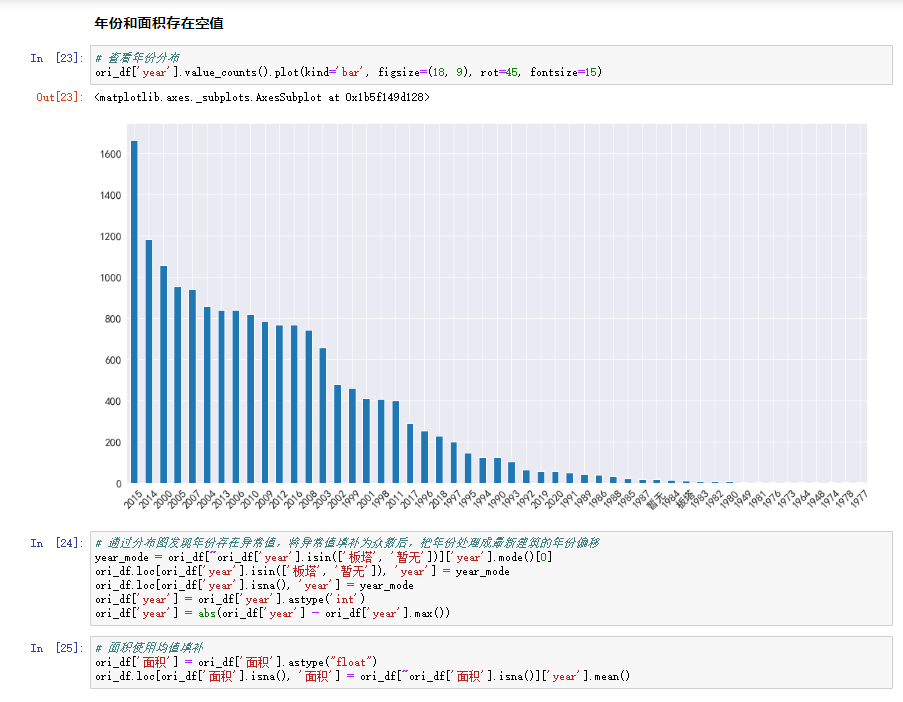


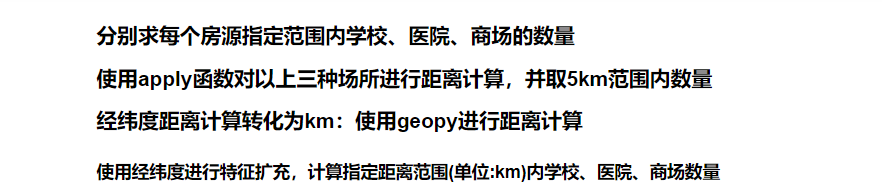






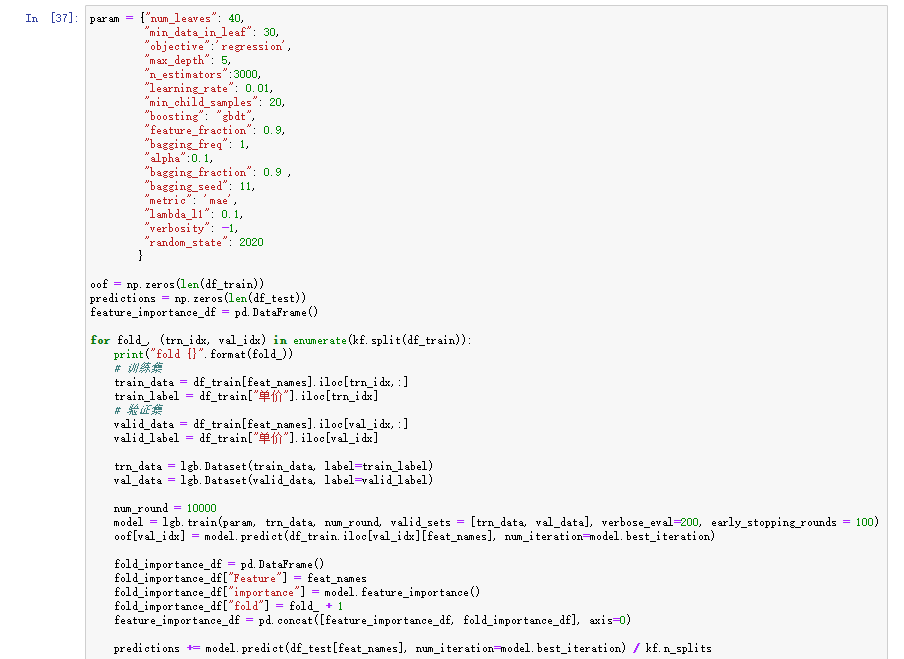


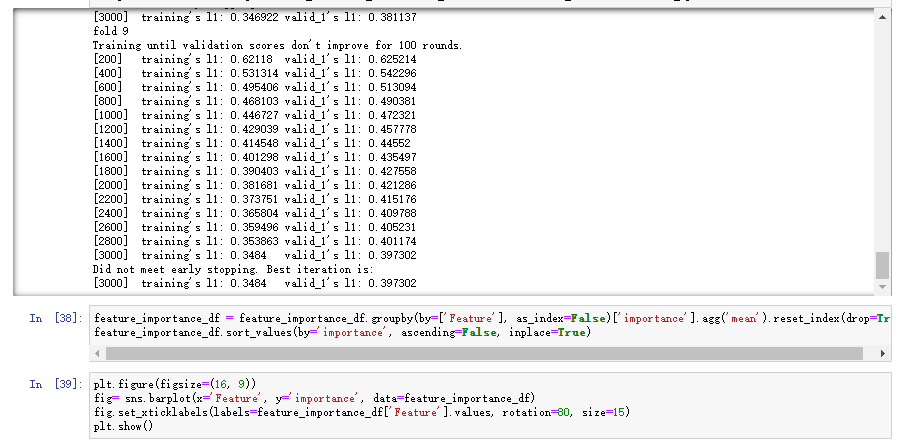


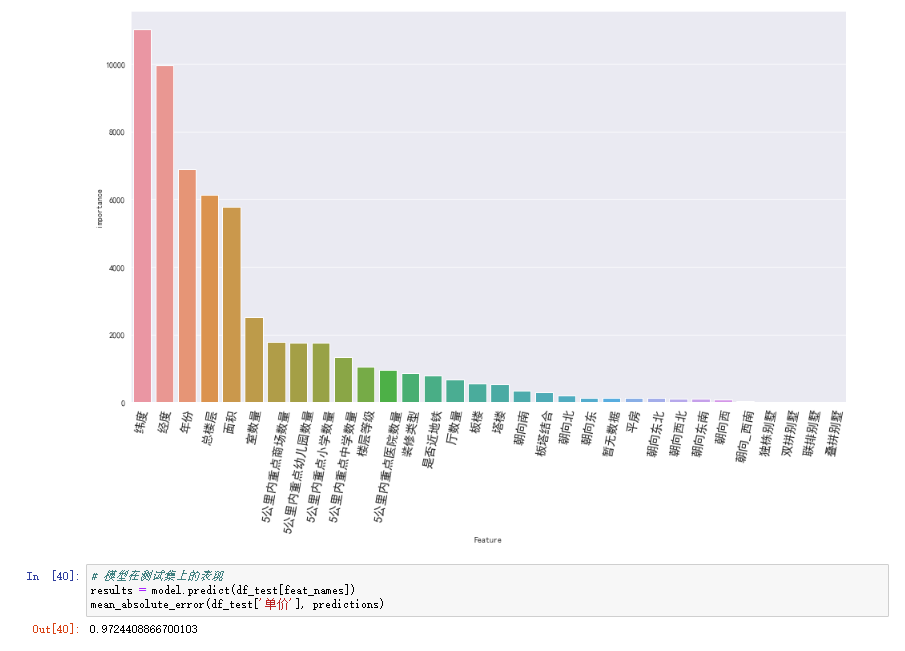


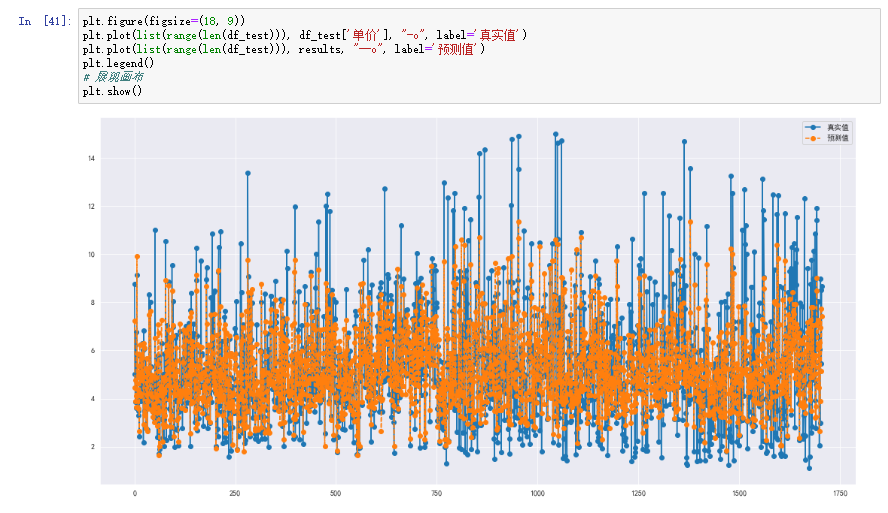


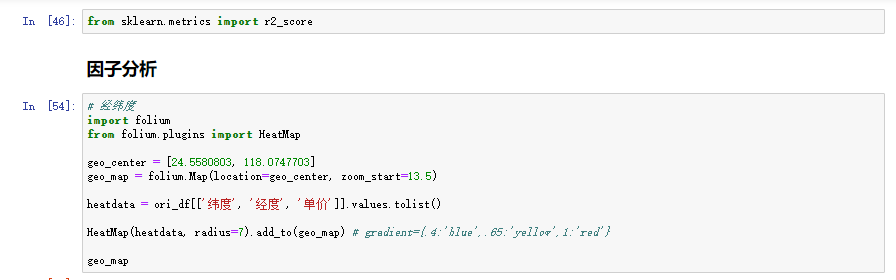


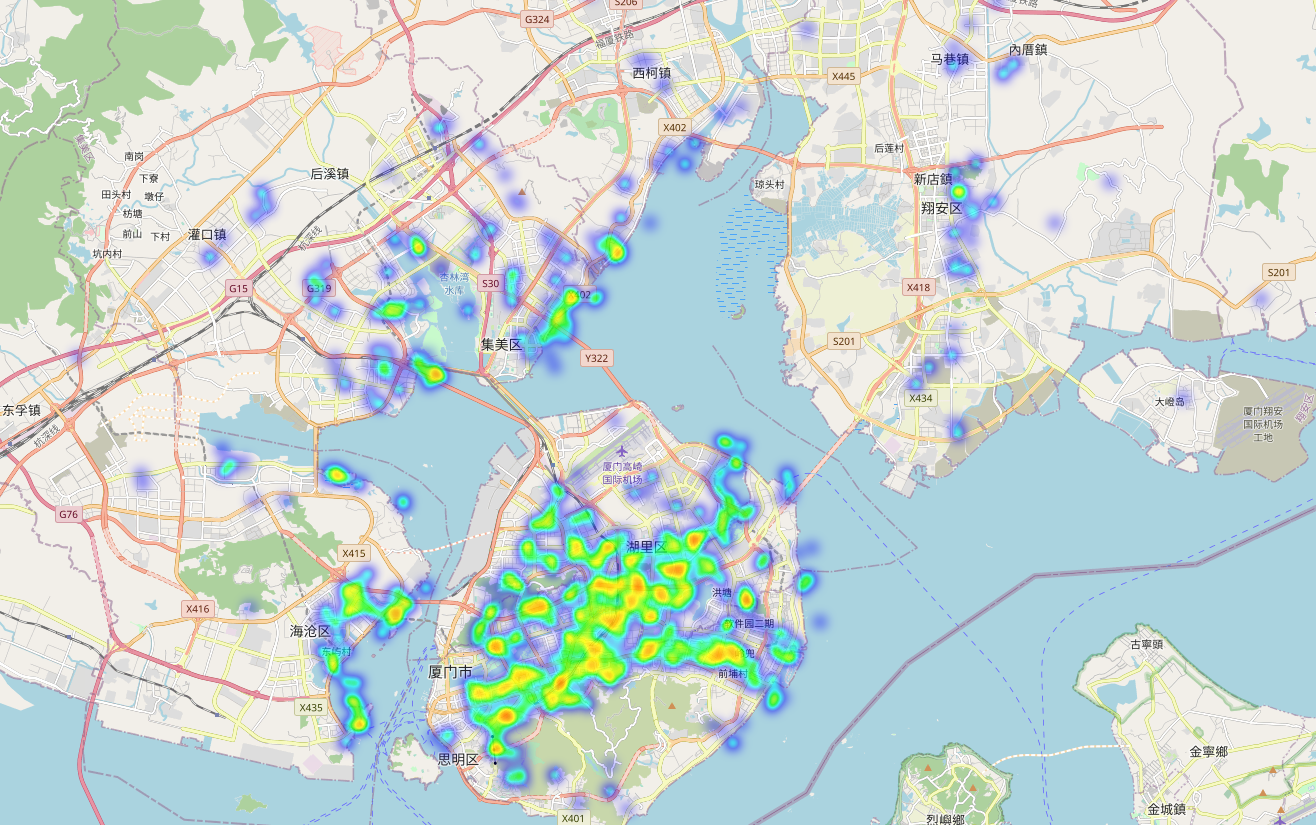


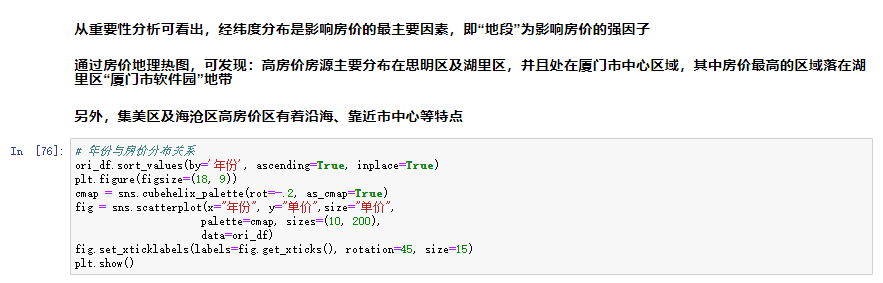


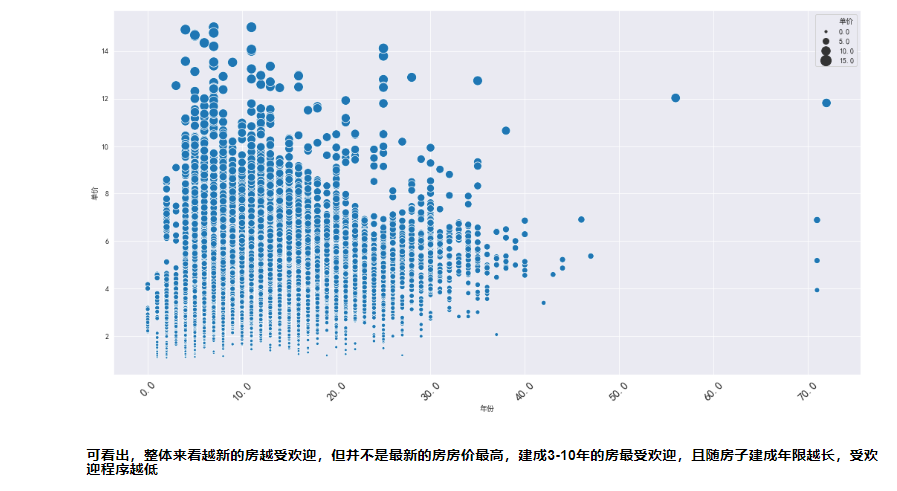


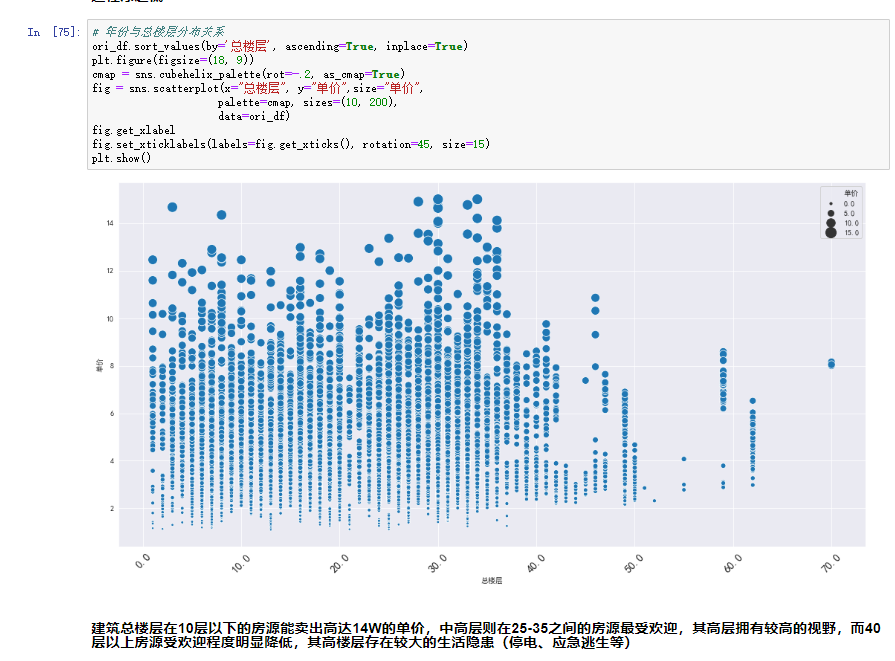


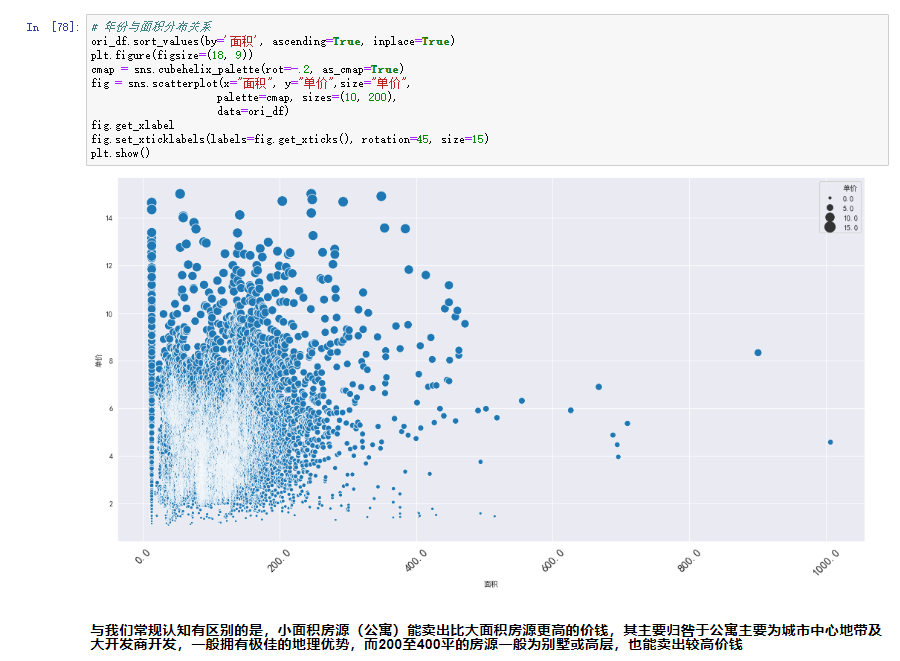


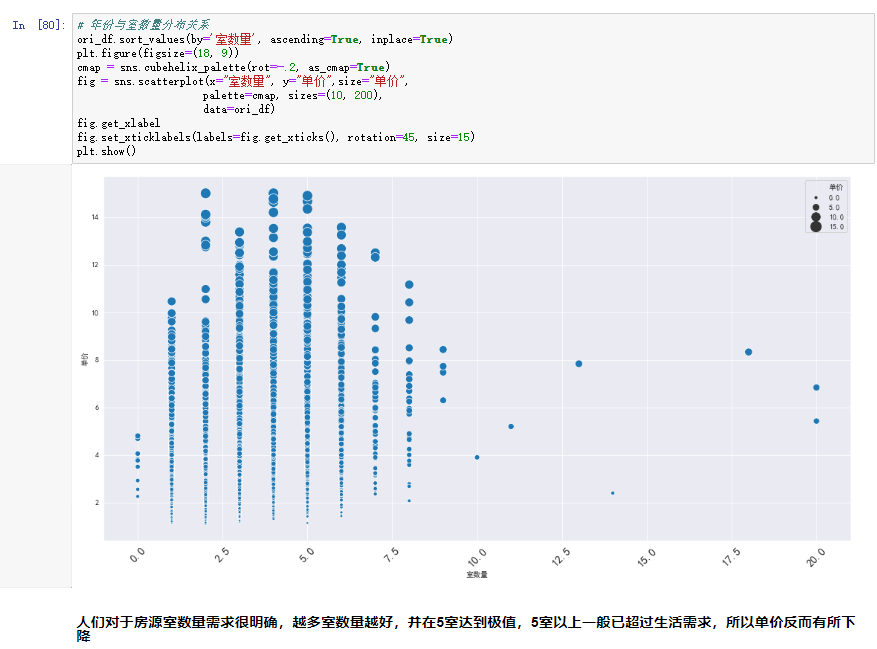












1. **结果的分析和讨论**
2. 从重要性分析可看出，经纬度分布是影响房价的最主要因素，即“地段”为影响房价的强因子
3. 通过房价地理热图，可发现：高房价房源主要分布在思明区及湖里区，并且处在厦门市中心区域，其中房价最高的区域落在湖里区“厦门市软件园”地带。另外，集美区及海沧区高房价区有着沿海、靠近市中心等特点
4. 从年份与房价的关系可看出，整体来看越新的房越受欢迎，但并不是最新的房房价最高，建成3-10年的房最受欢迎，且随房子建成年限越长，受欢迎程序越低。
5. 从年份与楼总的分布可以看出，建筑总楼层在10层以下的房源能卖出高达14W的单价，中高层则在25-35之间的房源最受欢迎，其高层拥有较高的视野，而40层以上房源受欢迎程度明显降低，其高楼层存在较大的生活隐患（停电、应急逃生等）。
6. 从年份与面积的关系可以看出，与我们常规认知有区别的是，小面积房源（公寓）能卖出比大面积房源更高的价钱，其主要归咎于公寓主要为城市中心地带及大开发商开发，一般拥有极佳的地理优势，而200至400平的房源一般为别墅或高层，也能卖出较高价钱。
7. 从年份与室的分布可以看出，人们对于房源室数量需求很明确，越多室数量越好，并在5室达到极值，5室以上一般已超过生活需求，所以单价反而有所下降。
8. **项目总结与未来展望**
9. **项目总结**

该项目通过爬虫爬取链家的数据，并进行存储，使用机器学习的方法预测房价并进行因子分析，总的来说是一个较为全方位的分析，但是此过程中机器学习的模型使用了lightgb模型，对于模型的使用和调参还可以更加完善一些。

1. **未来展望**

希望能有更加完善的分析思路，数据分析的内容能落地实用。