****

**房地产行业大数据科技报告**



**学 院 智能与计算学部**

**年 级 2021**

**姓 名 张可欣、黄琬婷、付凯**

**欣、马明朝、张瀚文**

**指导教师 王晓飞、仇超**

目 录

[第一章 绪论 1](#_Toc158672054)

[1.1 简要结论 1](#_Toc158672055)

[第二章 技术研究 2](#_Toc158672056)

[2.1 数据采集和清洗 2](#_Toc158672057)

[2.1.1 数据采集步骤 2](#_Toc158672058)

[2.1.2 数据清洗实现步骤 2](#_Toc158672059)

[2.1.3 数据清洗部分代码展示 3](#_Toc158672060)

[2.2 统计与可视化（黄琬婷） 3](#_Toc158672061)

[2.2.1 可视化技术 3](#_Toc158672062)

[2.2.2 房屋与地理位置可视化分析 4](#_Toc158672063)

[2.2.3 房屋与自身影响因子可视化分析 10](#_Toc158672064)

[2.2.4 房屋与外界影响因素可视化分析 15](#_Toc158672065)

[2.3 大数据分析算法——MapReduce（付凯欣） 17](#_Toc158672066)

[2.3.1 原理及选取目的 17](#_Toc158672067)

[2.3.2 步骤 18](#_Toc158672068)

[2.3.3 结果与分析 19](#_Toc158672069)

[2.4 大数据分析算法——K-means（付凯欣） 20](#_Toc158672070)

[2.4.1 原理及选取目的 20](#_Toc158672071)

[2.4.2 步骤 20](#_Toc158672072)

[2.4.3 结果与分析 21](#_Toc158672073)

[2.5 大数据分析算法——决策树（张可欣） 22](#_Toc158672074)

[2.5.1 原理及选取目的 22](#_Toc158672075)

[2.5.2 步骤 23](#_Toc158672076)

[2.5.3 结果和分析 24](#_Toc158672077)

[第三章 总结 27](#_Toc158672078)

[3.1 整体总结 27](#_Toc158672079)

[3.2 反思与展望 27](#_Toc158672080)

[附录 28](#_Toc158672081)

# 绪论

本项目报告旨在透过大数据工具与分析方法，深入研究后疫情时代背景下的天津市房地产市场现状。考虑到当前购房者和租客面临的经济形势，本研究聚焦于二手房与租赁市场，借助大数据分析技术，对天津市房地产市场进行全面而深入的量化分析。我们希望通过对二手房和租房市场价格及其决定性因素（如房屋面积、位置和户型等）的研究，揭示二手房市场的内在运作机制。以此为潜在买家和租客提供市场洞察，同时能够为房产业内的政策制定者和市场参与者提供数据支撑和决策依据。

## 1.1 简要结论

1. 二手房市场在面积和价格外最受关注的房屋特征为房屋朝向与采光，其次为有无电梯与绿化率。

2.租房者最关注的房屋特征为装修和家具家电配备情况，其次为地理位置与交通方便性。

3.房屋所处的城市区位与房屋面积是构成房屋价格的最关键因素。

# 技术研究

## 2.1 数据采集和清洗

### 2.1.1 数据采集步骤



图2-1.1 数据采集步骤图

### 2.1.2 数据清洗实现步骤

1. 导入必要的库。

2. 加载数据：使用Pandas库读取数据文件，将数据加载到DataFrame中。

3. 初步数据探索：查看数据的前几行以及数据的基本信息，以了解数据的结构、特征和潜在问题。

4. 处理缺失值。

删除包含缺失值的行

填充缺失值

5. 处理重复值。

6.处理异常值。

根据阈值过滤异常值

7.文本数据清洗：如果数据包含文本信息，可能需要进行文本数据清洗，包括去除特殊字符、停用词、标点符号等操作。

8.数据转换和整合。

数据类型转换

合并数据集

创建新特征

9. 数据验证和测试。

### 2.1.3 数据清洗部分代码展示

这里因为格式要求，保存的编码格式为utf-8-sig，在处理完之后，数据删除了存在空单元格或重复的数据。



图2-1.2 数据采集部分代码

## 2.2 统计与可视化（黄琬婷）

### 2.2.1 可视化技术

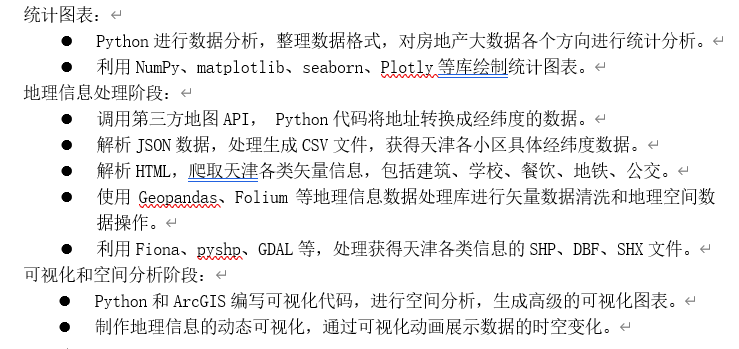


图2-2 统计图表技术图

### 2.2.2 房屋与地理位置可视化分析

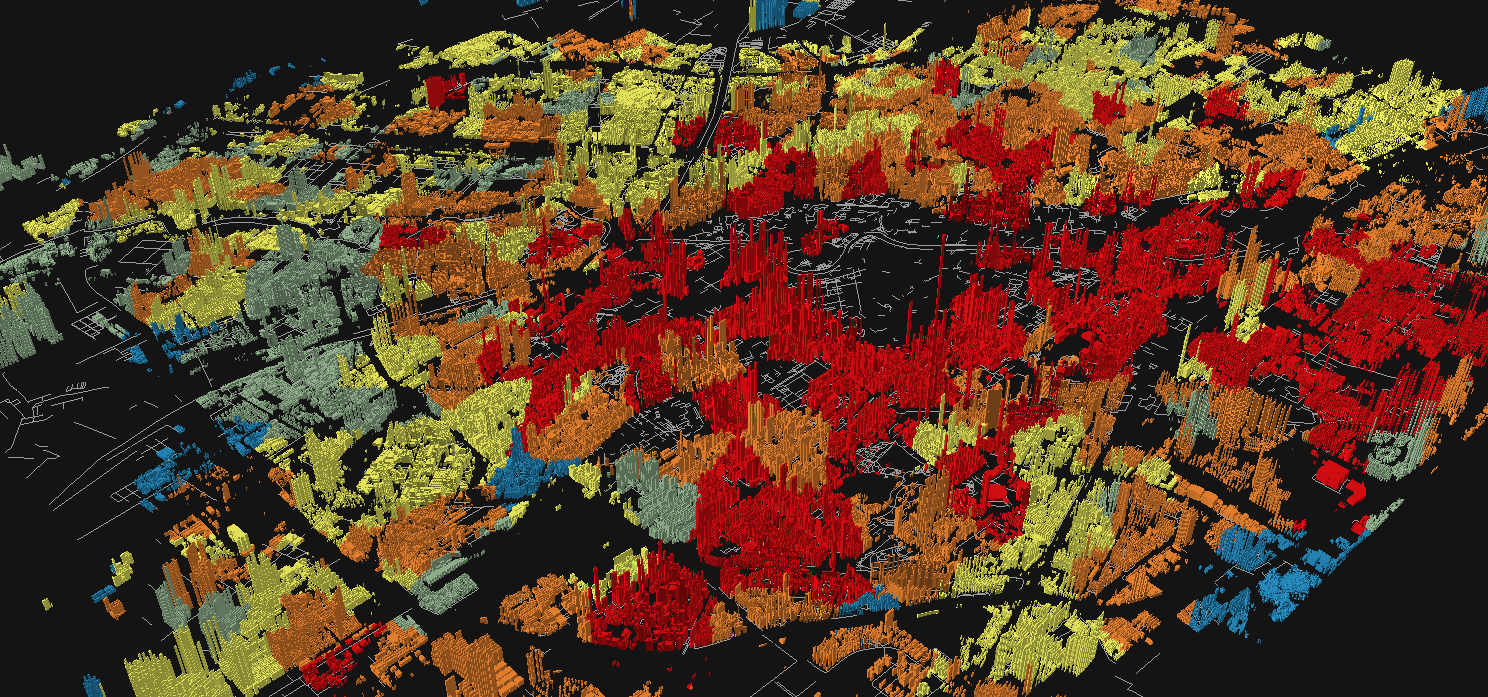


图2-3 天津建筑热度3D图

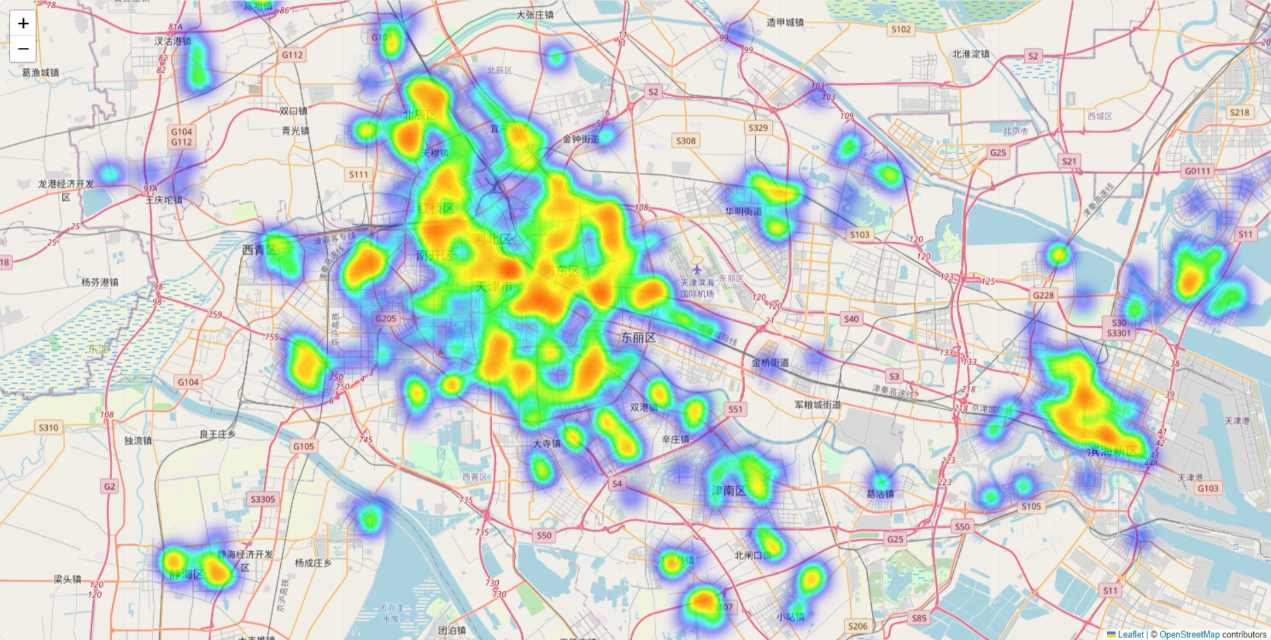
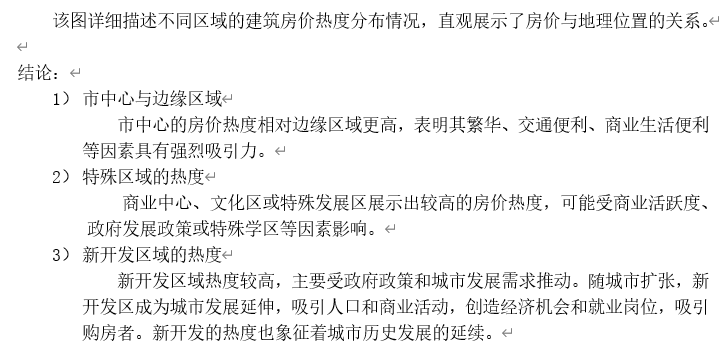


图2-4 天津热度图



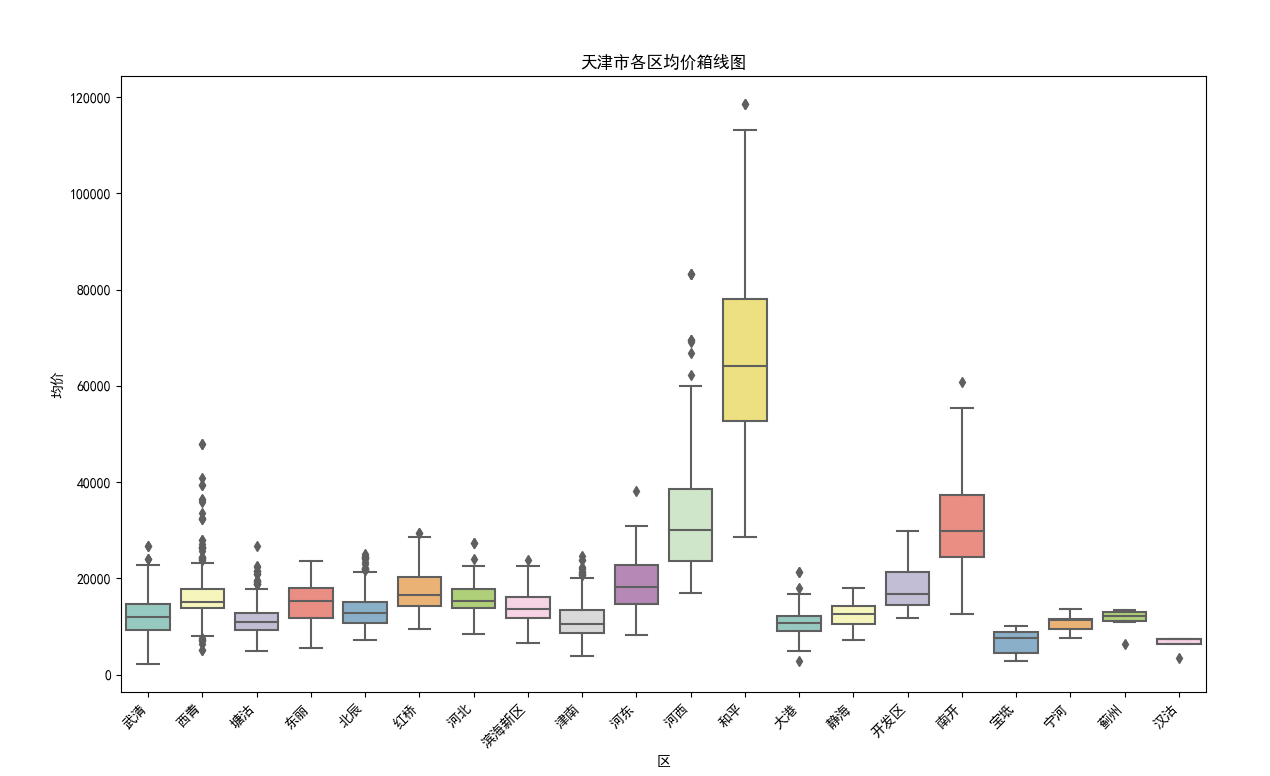


图2-5 天津各区均价箱线图

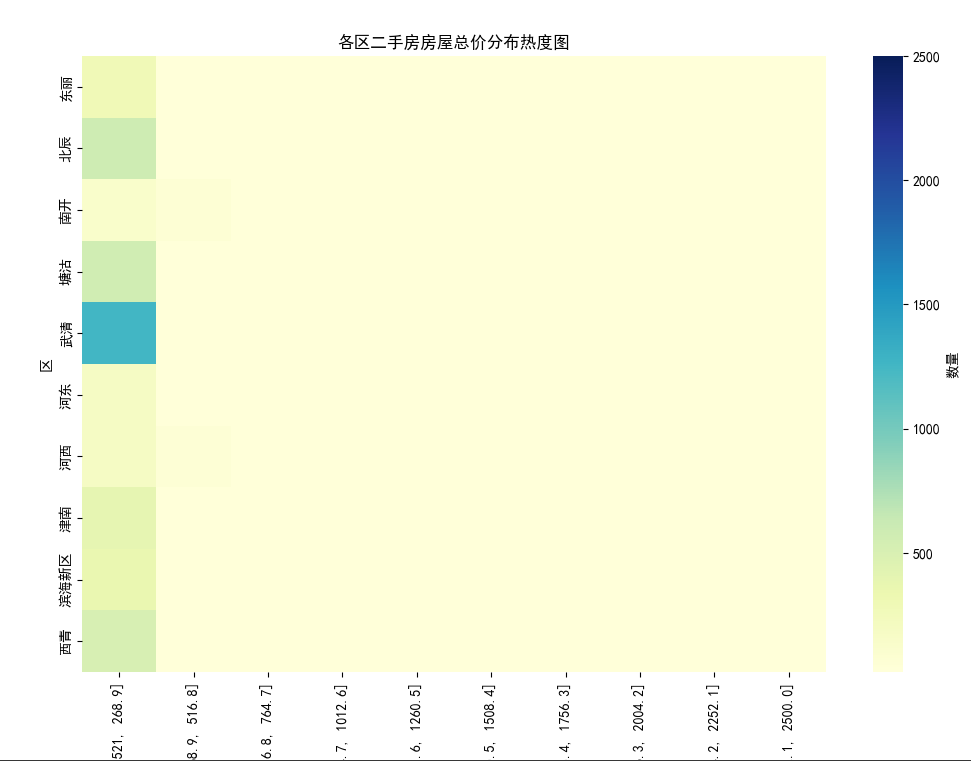
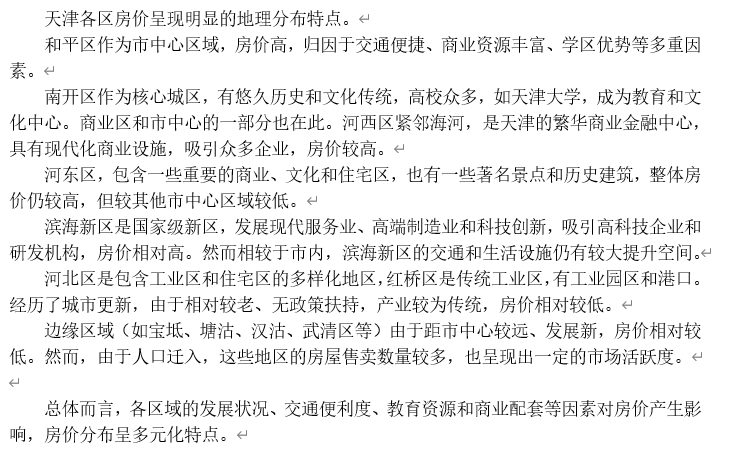


图2-6 天津各区二手房房屋总价分布热度图



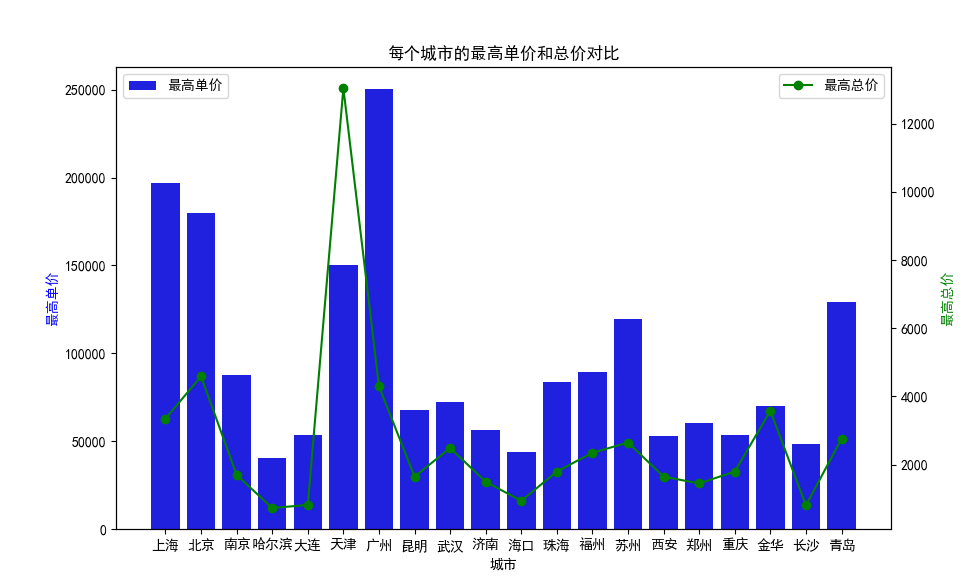


图2-7 不同城市房屋最高单价和总价对比图

比较多个城市房屋价格，发现天津最高总价非常高。具体而言，这些房子位于天津中心历史悠久的区域，开发早，配套设施完备，医教商发达，生活交通便利。供应有限，属于市场上的稀缺资源，具有不可替代性。同时，部分房子位于具有独特文化和历史价值的区域，如古文化遗产区，房价受特殊价值影响

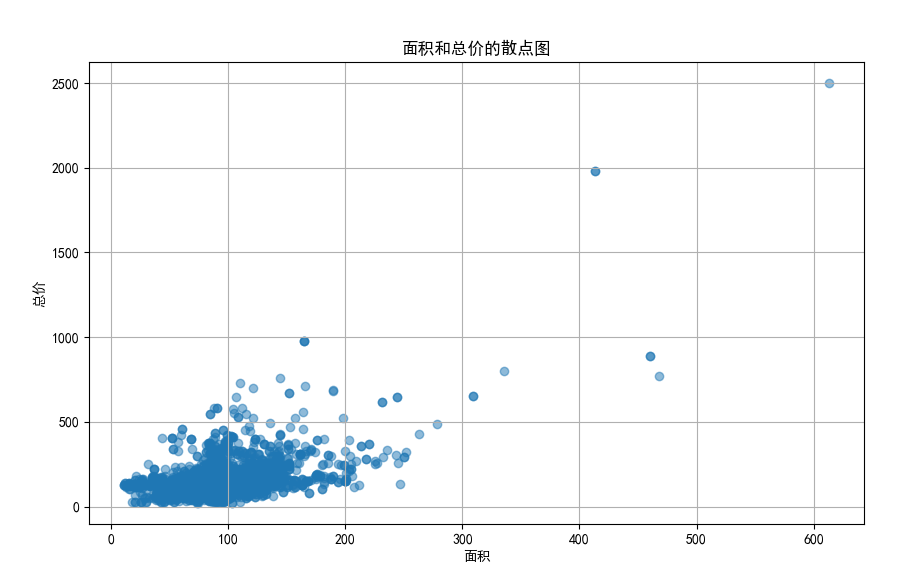


图2-8 天津面积与总价散点图

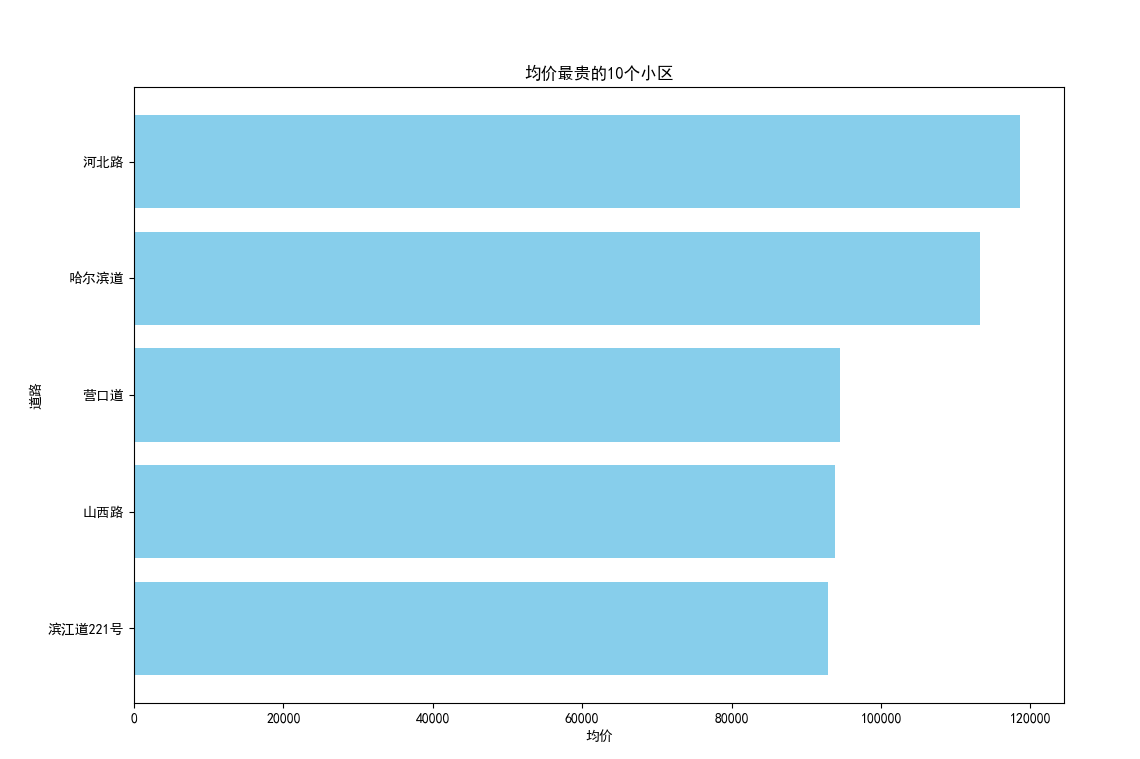


图2-9 天津均价最贵的小区图

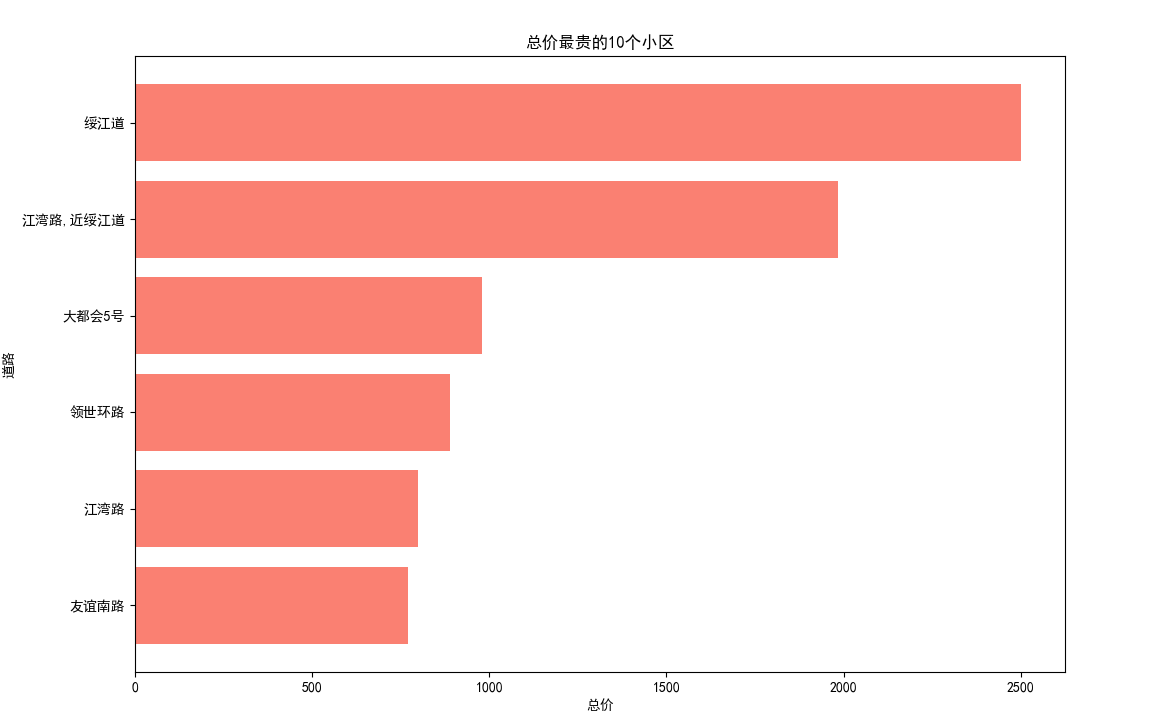


图2-10 天津总价最贵的小区图

图2-8显示天津房价主要集中在200平方米以内、总价500万以内，符合大多数家庭需求，反映房地产市场的普遍趋势。然而，也存在个别高价房产，可能是豪宅或特殊类型房产。

图2-9显示均价最高的地区主要位于和平区核心地段，拥有丰富商业和医教资源，同时是文化遗产所在地，使得该区域均价最高。

图2-10显示总价最贵的地区主要分布在和平区核心地段、津南和西青区的新开发别墅地段。这些地区环境优美，拥有大型购物商业中心，具有宜居特点。

结论：

1） 分层特征

天津房地产市场呈现分层特征，部分区域房价相对较高，吸引更富裕购房者。这也反映了城市内不同地段的不同发展水平和特色。

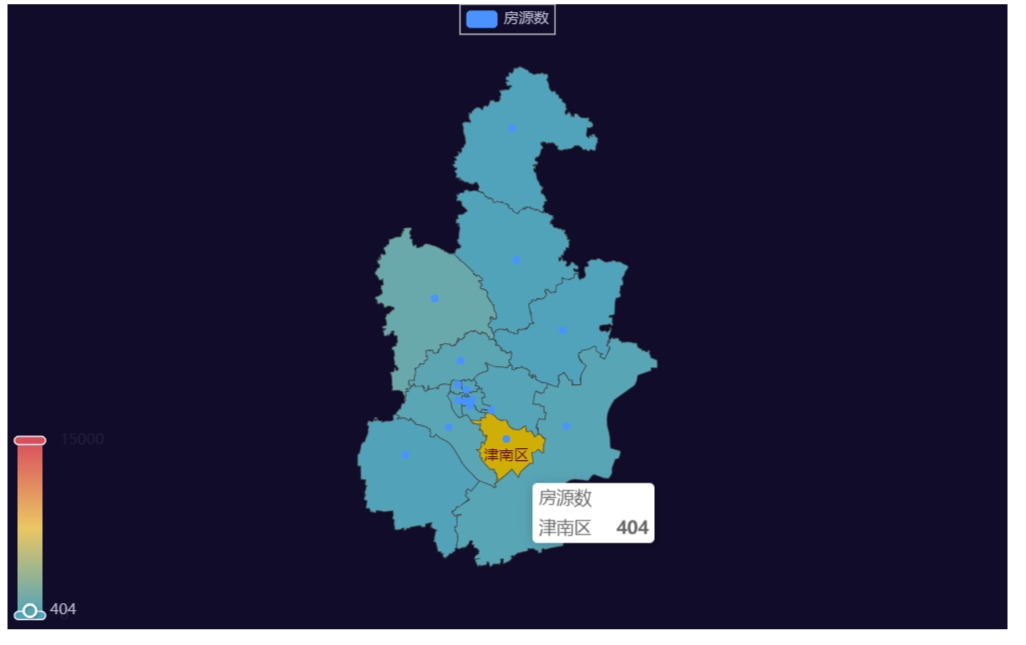
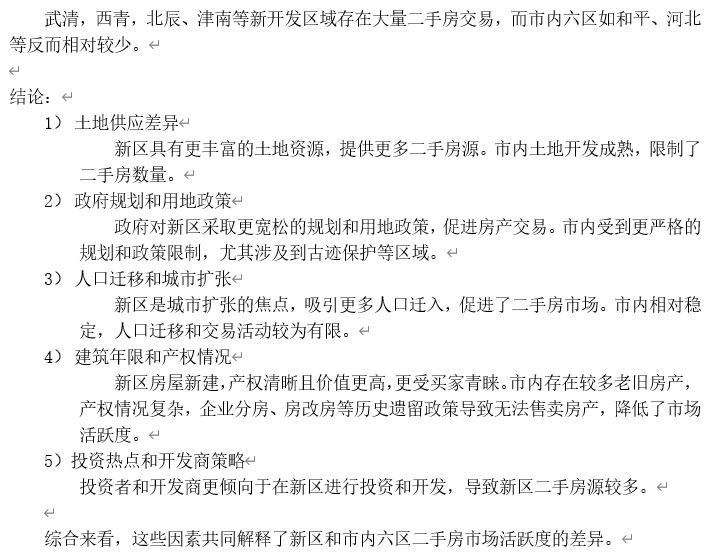


图2-11 天津各区房源数量图



### 2.2.3 房屋与自身影响因子可视化分析

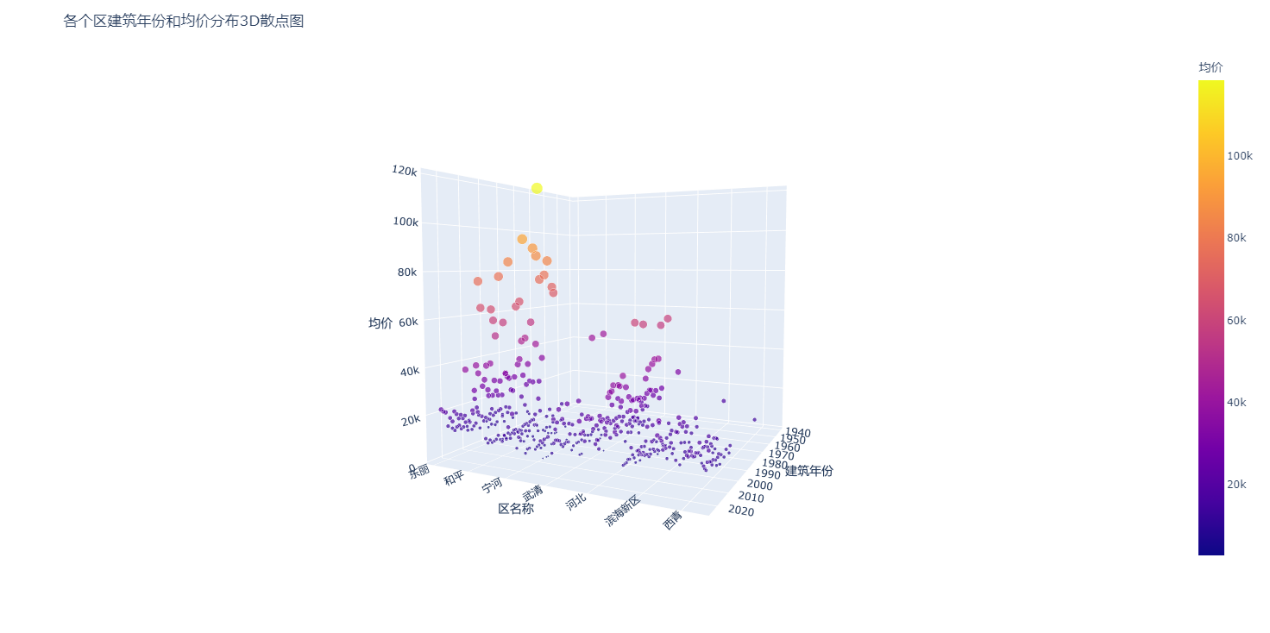


图2-12 天津各区建筑年份与均价分布3D散点图

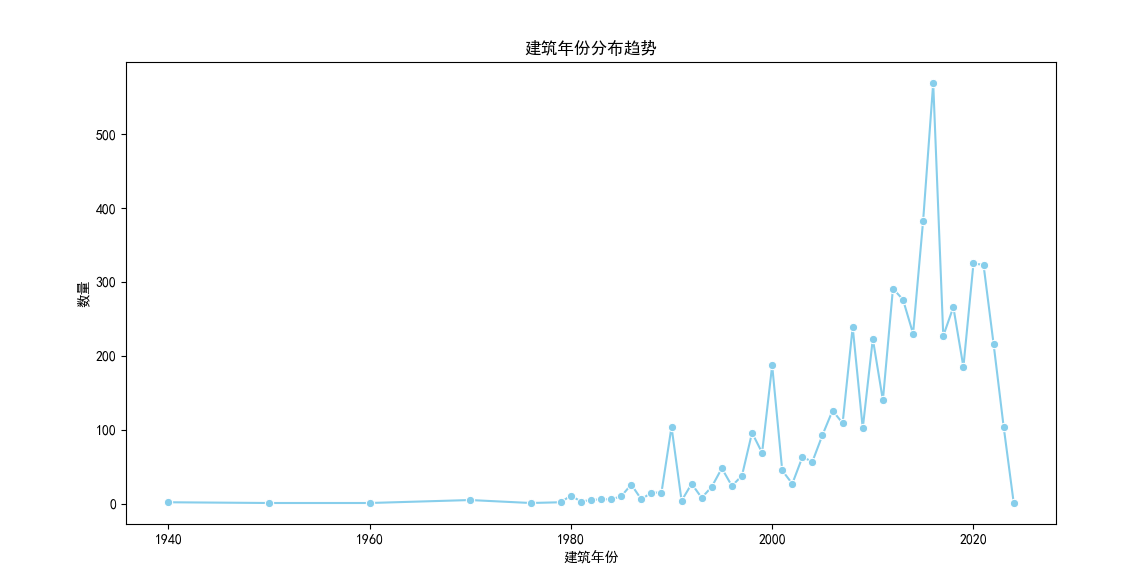
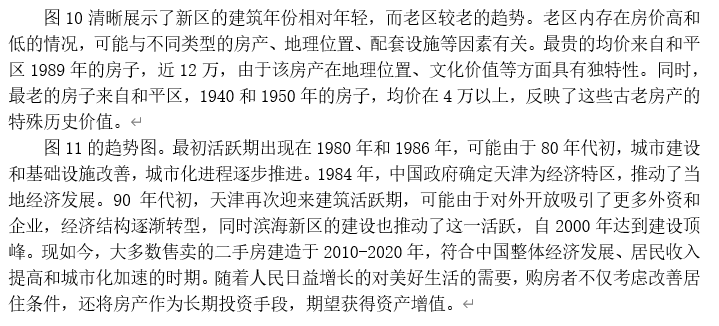
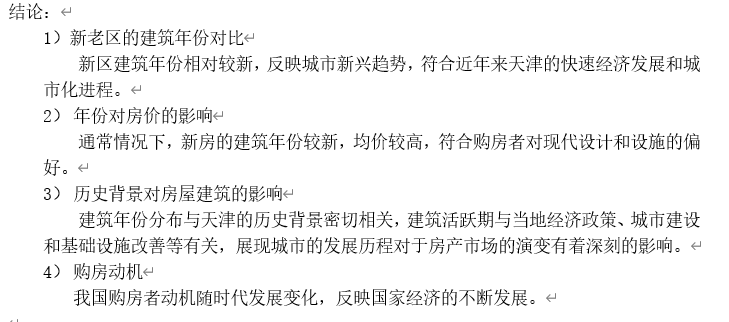


图2-13 天津建筑年份分布趋势





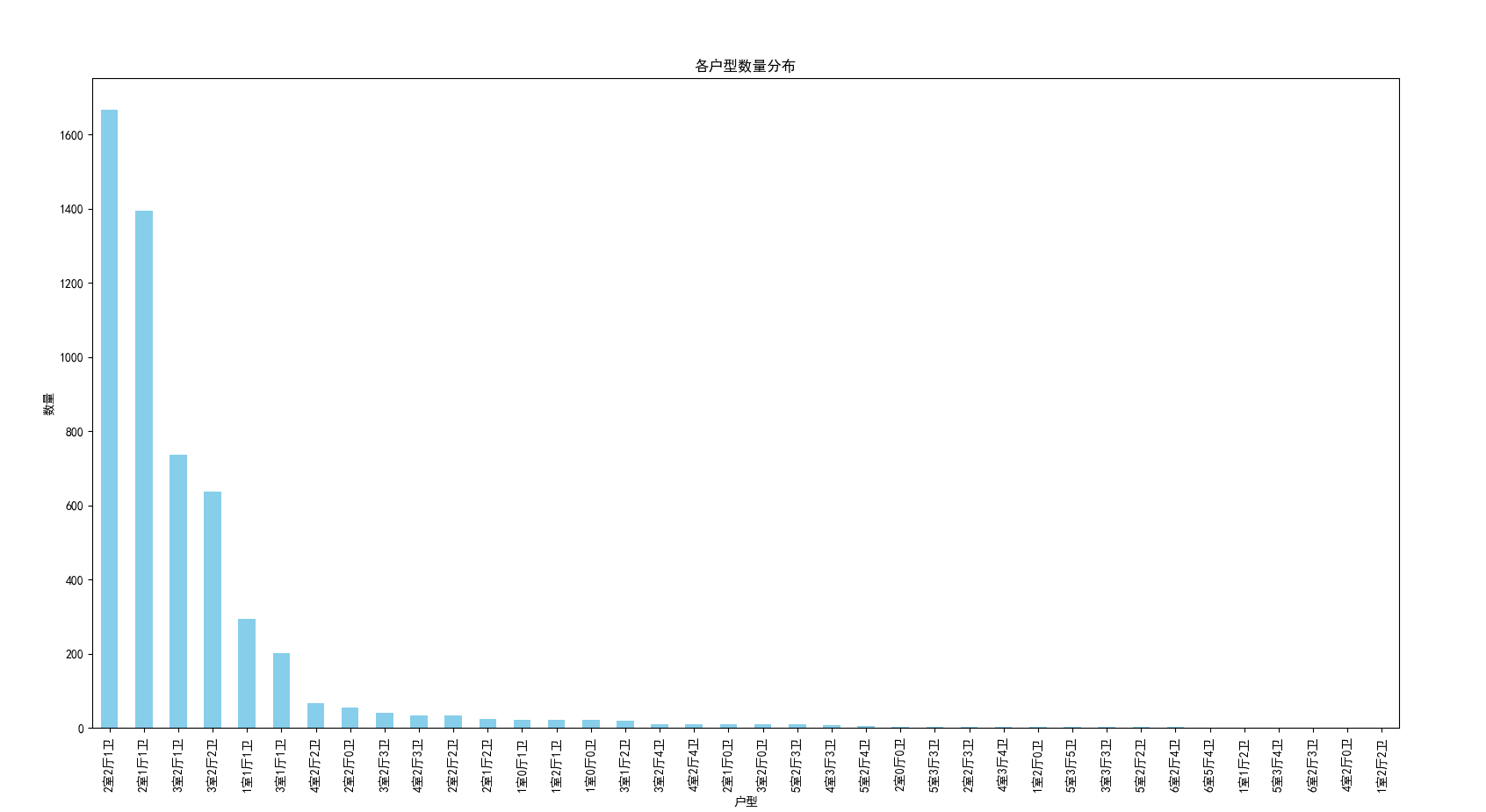


图2-14 天津各户型数量分布

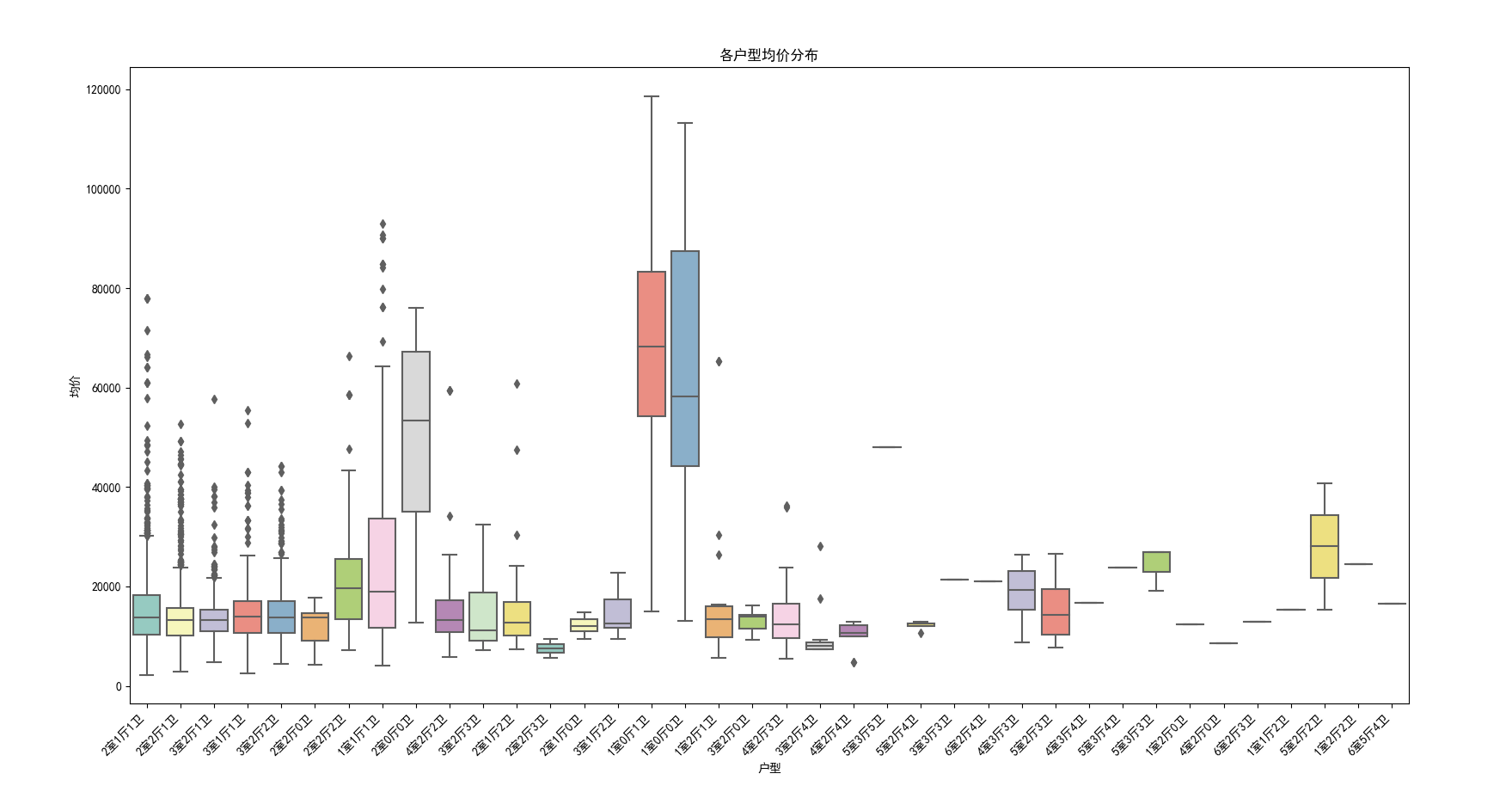


图2-15 天津各户型均价

天津户型主要以两室两厅一卫和两室一厅一卫为主，其他户型的数量逐渐减少。这说明大多数的房子设计适合一家人一起居住，符合家庭生活的需求。

各户型的均价情况中，一室、两室或一室一卫的房子显示出较高的均价，主要因为位于热门学区，独特的地理位置加特殊的房屋设计，满足了许多人需求，从而均价提高。

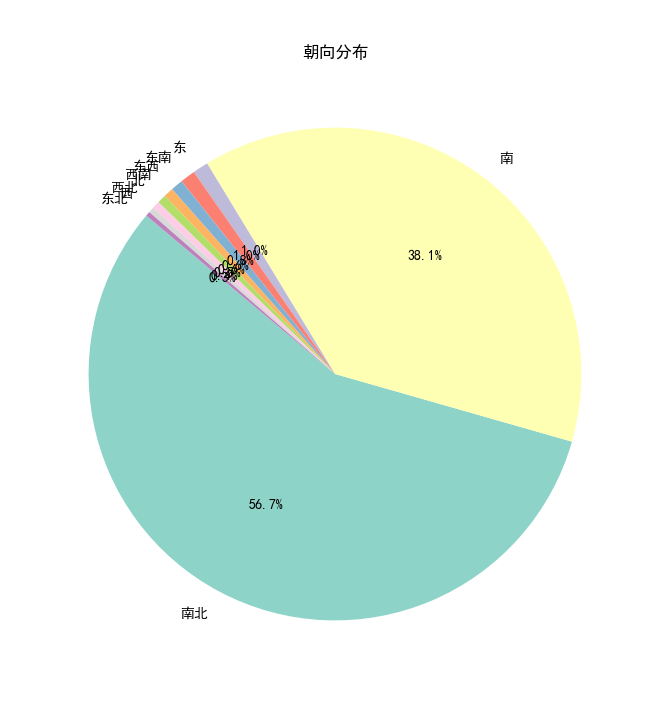


图2-16 天津房屋朝向分布

天津房屋主要以南北朝向为主，其次是南朝向，南北+南朝向的组合占据了95%左右。这种朝向分布符合北方气候，冬季可获得最大限度阳光，提高采暖效果。

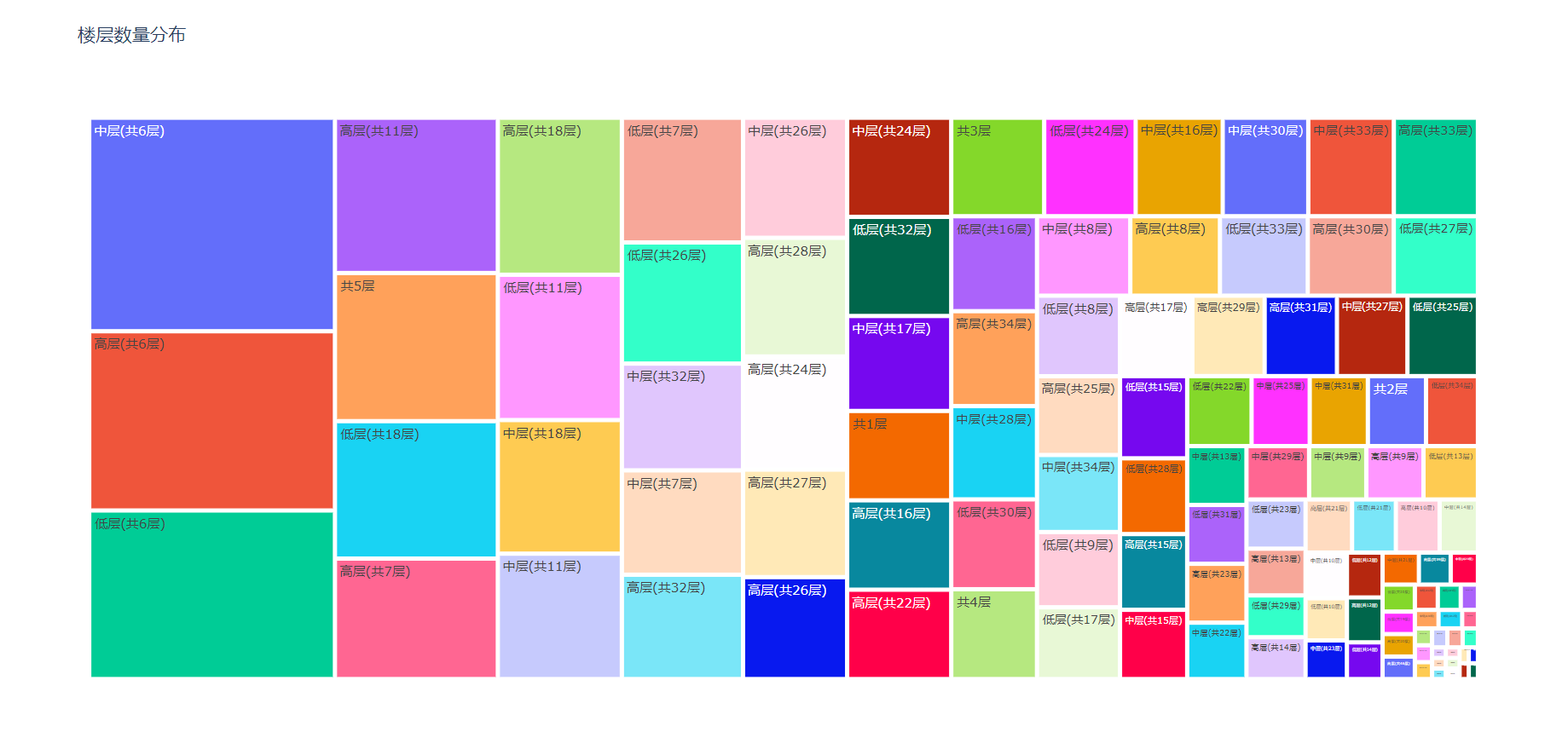


图2-17 天津楼层分布

天津房屋楼层主要集中在6层左右的洋房。

结论：

1） 地理条件

天津地理位置特殊，靠近海河、地势平坦。较低的楼层更容易与自然地形融合，减少地基工程难度，降低建设成本，提高施工效率。这使得在城市规划中更容易采用低层建筑，而不需要过多考虑地形起伏对建筑的影响。

2） 城市规划

天津作为具有悠久历史和文化传统的城市，在建筑风格和高度上有传统习惯。较低的楼层高度符合当地的历史建筑风格和文化特点，保留古建筑。

3） 家庭生活习惯

在中华传统文化之中，低层建筑更符合传统的家庭结构和居民对自然环境、社区生活的期望。

### 2.2.4 房屋与外界影响因素可视化分析

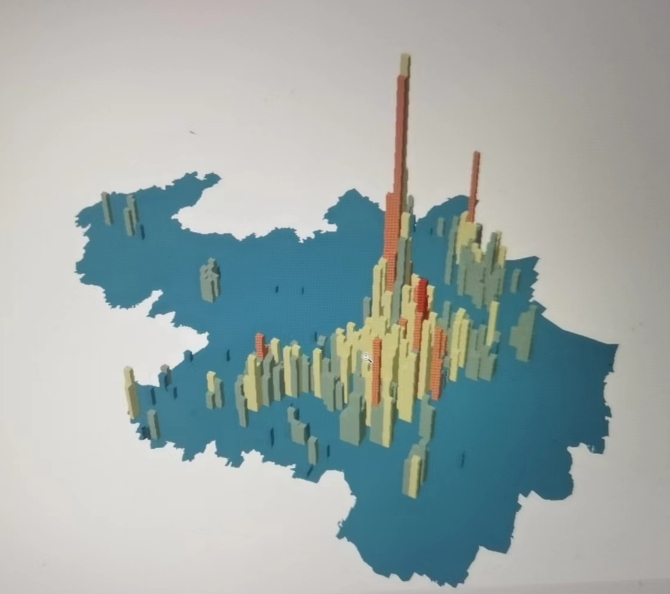


图2-18 天津学校与房价

如图2-18，学区好坏显著影响周边房价。好学区吸引更多家庭购房，因为希望为孩子提供更好教育资源，导致房价上升。此外，好学区购房是一种长期投资，因为该房产未来更容易升值，这预期投资回报也推动学区房价格上涨。

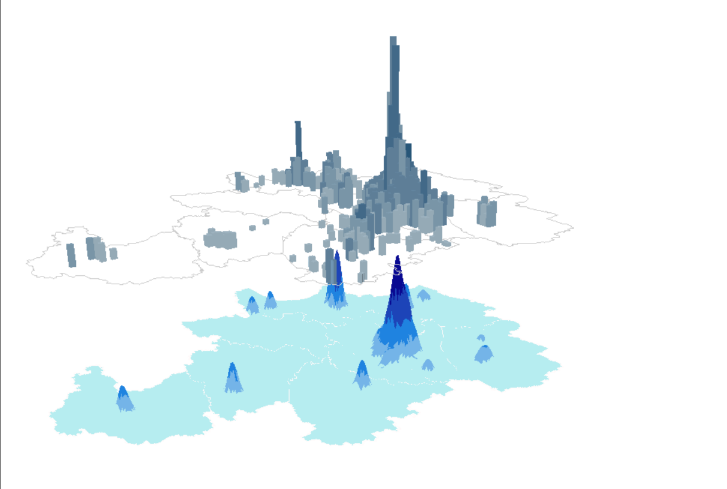


图2-19 餐饮行业与房价分布图

图2-19，餐饮业繁荣的地段通常表现为房价相对较高的区域。

结论：

1） 人流集中与商业活跃度

餐饮繁荣吸引大量顾客，导致人流聚集，商铺和娱乐场所也可能在此设立。

2） 交通便利与生活便利性

该地段通常有发达的交通网络，使人们更容易前来就餐和参与商业活动，提高生活便利性，包括购物、娱乐等各类服务。

3） 商业区的综合影响

可能是城市商业核心区，对该区域整体经济和社会发展有积极影响，推动周边房价上涨。

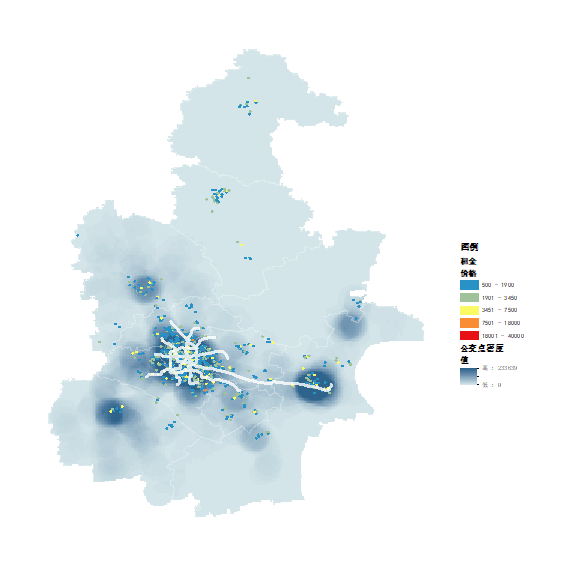


图2-20 公交地铁与房价分布图

如图2-20，公交站线与地铁线覆盖的区域通常表现出租房房价相对较高的趋势。

结论：

1） 交通便利与房价正相关

交通便利性直接关系到居民的生活便利性，更受租房者青睐，推动了这些区域的房价上涨。

2） 交通网络密集区域的优势

公交地铁网络密集的地方可能形成了交通节点，周边设施、商业区等更加发达。租房者倾向于选择这些区域，使得这些地段的租房房价较高。

## 2.3 大数据分析算法——MapReduce（付凯欣）

### 2.3.1 原理及选取目的

MapReduce算法用于大规模数据集的并行计算，包括Map阶段和Reduce阶段。在Map阶段，原始数据被分块并并行处理，转换为键值对的形式；Reduce阶段对键相同的值进行合并和计算，生成最终结果。实验使用存储在HDFS中数据，编写Hadoop的MapReduce任务，希望通过MapReduce算法对购房和租房热词数据进行词频统计，了解市场热点趋势及消费者关注因素，以关注天津市房价市场需求。

### 2.3.2 步骤

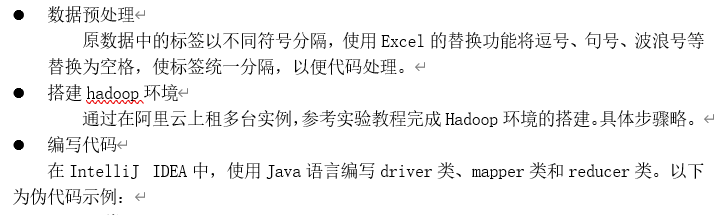


图2-21 运行前准备

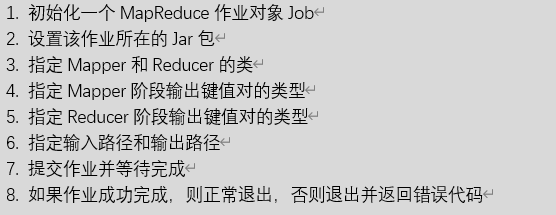


图2-22 Diver类

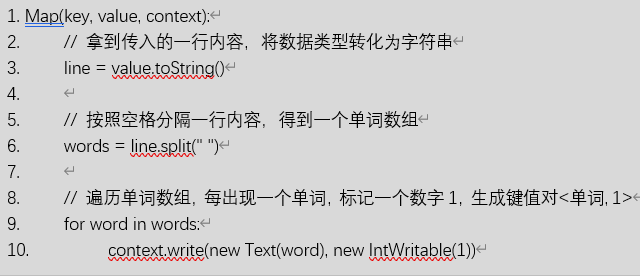


图2-23 Mapper类

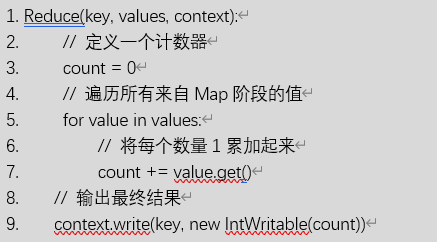


图2-24 Reducer类

在IDEA中使用package功能将代码打包成jar包，即MapReduce\_WordCount.jar，然后上传至搭建好Hadoop环境的服务器。

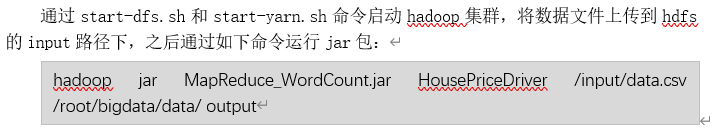


图2-25 运行实现方法

### 2.3.3 结果与分析



图2-26 二手房词云

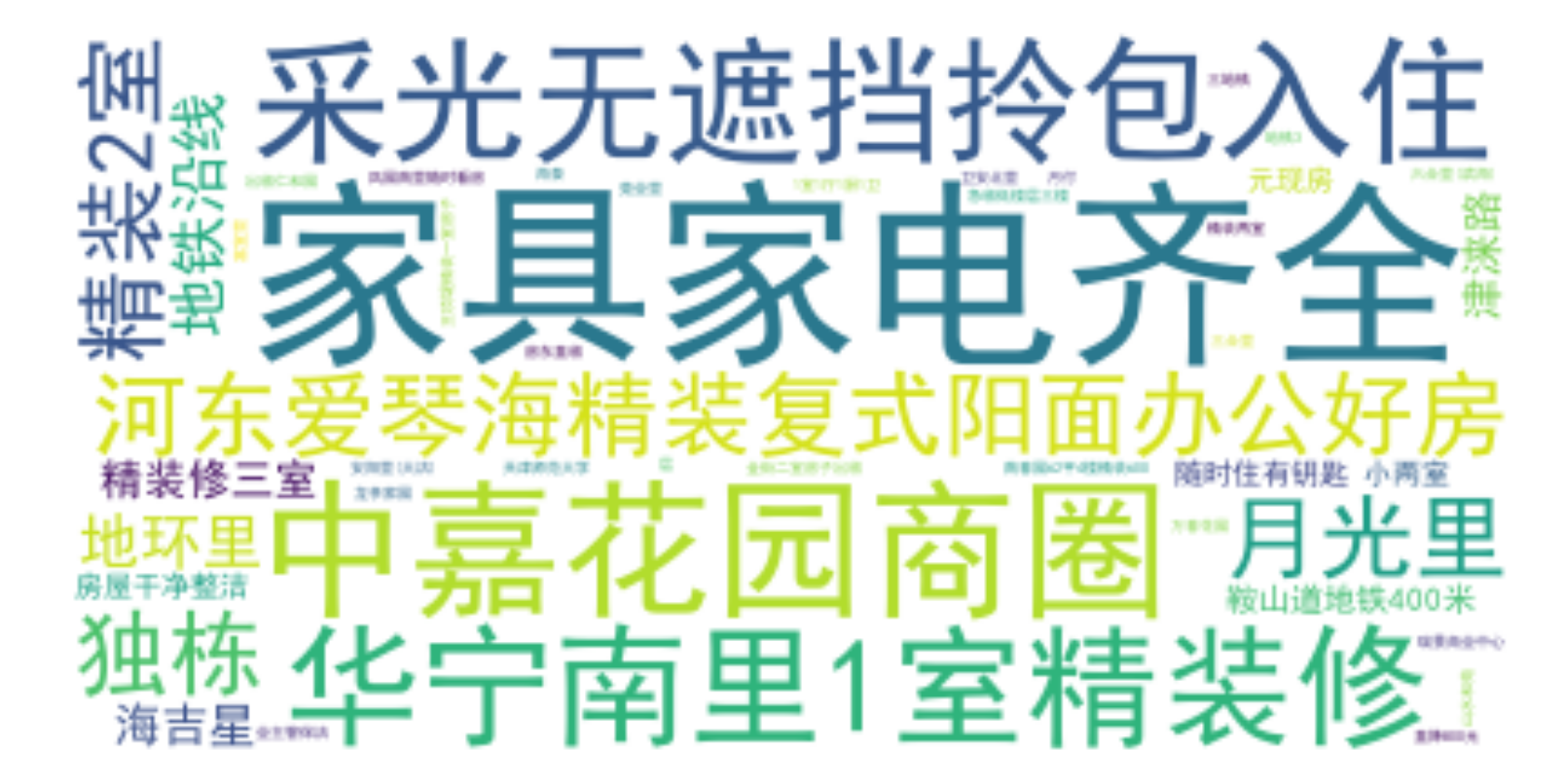


图2-27 租房词云

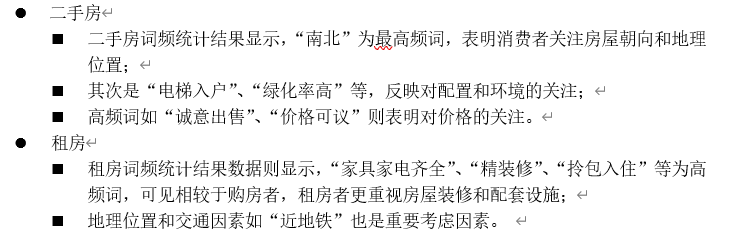


图2-28 MapReduce分析与结论

## 2.4 大数据分析算法——K-means（付凯欣）

### 2.4.1 原理及选取目的

K-means是一种无监督机器学习算法，用于聚类分析。聚类分析有助于发现数据中隐藏的模式和结构。在二手房市场，可能存在不同价位段、特征相似的房源集群，通过聚类可以揭示这些潜在的分布规律，了解不同聚类中房源的特征，有助于购房者更明智地选择符合其需求的房产。

### 2.4.2 步骤

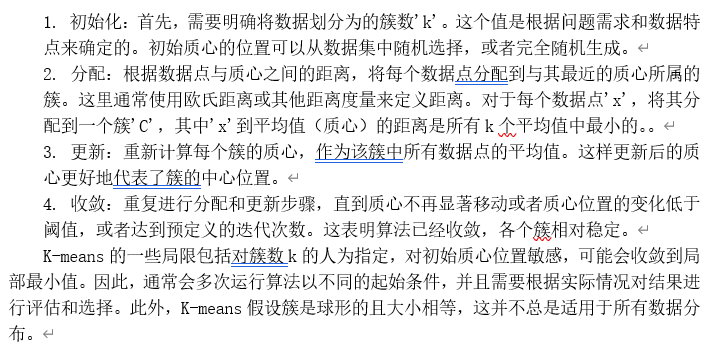


图2-29 K-means步骤图

### 2.4.3 结果与分析

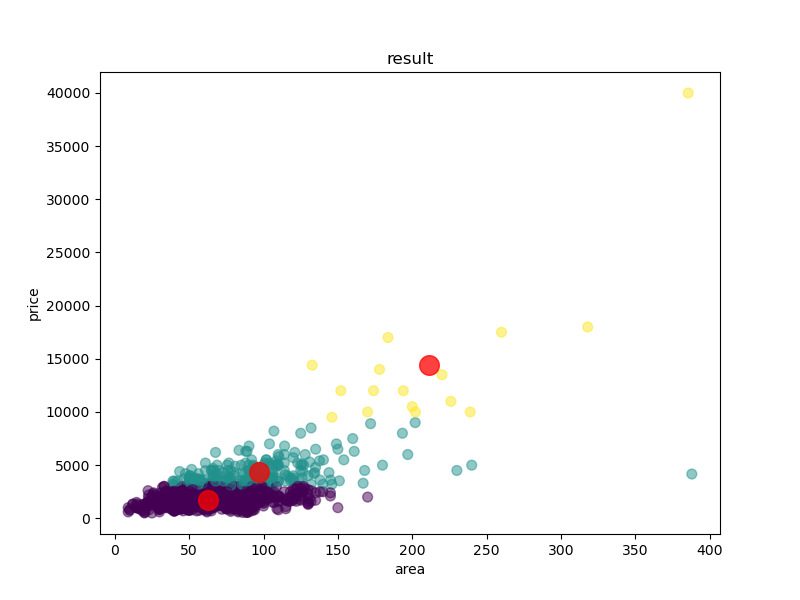


图2-29 单价与面积K-means聚类分析图

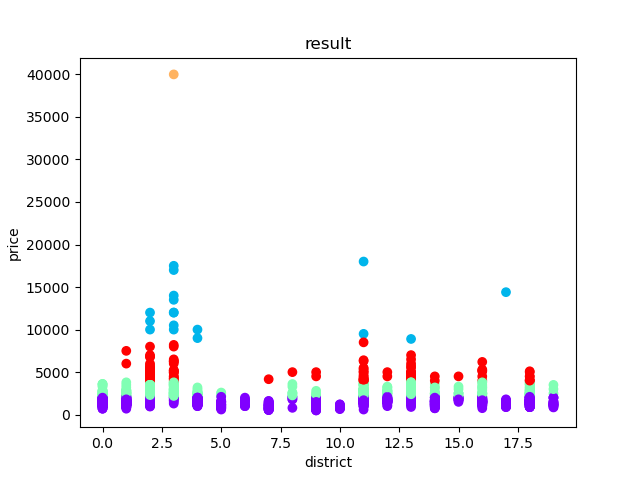


图2-30 单价与地理位置聚类分析图

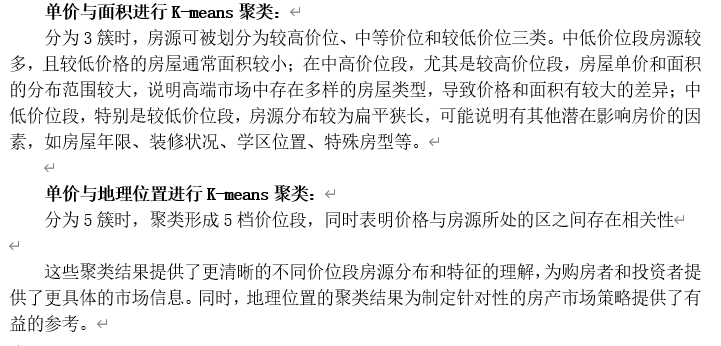


图2-31 K-means总结分析

## 2.5 大数据分析算法——决策树（张可欣）

### 2.5.1 原理及选取目的

决策树预测分类在房价大数据分析中具有关键作用，其优势包括理解关键特征、建立可解释的模型、捕捉非线性关系，处理混合类型数据，并对噪声和异常值具有较强鲁棒性。由于房价通常受自身、外界、市场波动等多种因素的综合影响、数据规模大且含有多种类型，决策树成为房价预测领域中常用的建模工具之一。

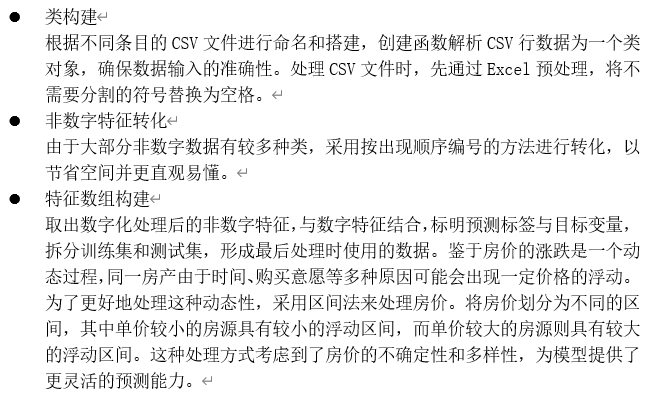


图2-32 数据处理原理图

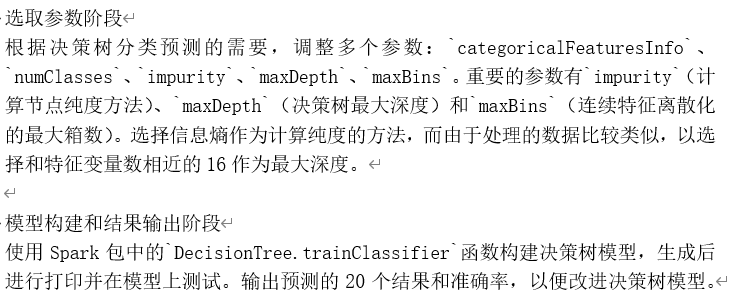


图2-33 选取参数和模型构建原理图

### 2.5.2 步骤

对房产大数据，使用scala语言编写的spark决策树分类预测进行分析。以下为代码流程

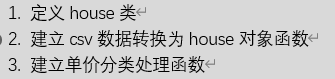


图2-34 非main函数部分

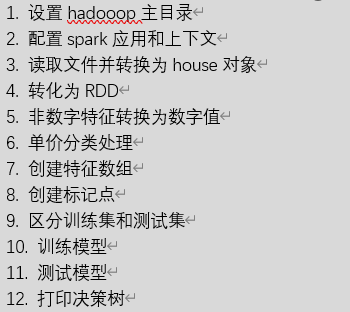


图2-35 main函数部分

使用start-dfs.sh和start-yarn.sh命令启动hadoop集群，构建sbt环境，将代码文件和数据文件等按以下顺序分布，以下均居于运行文件根目录：

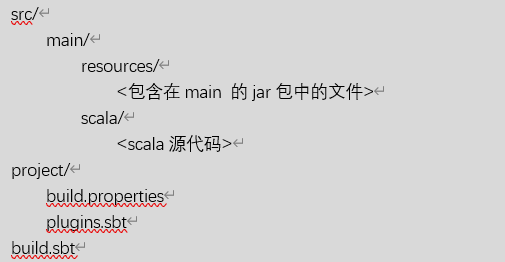


图2-36 文件结构图

之后该文件中使用sbt run运行，既可自动配置环境和编译。

### 2.5.3 结果和分析

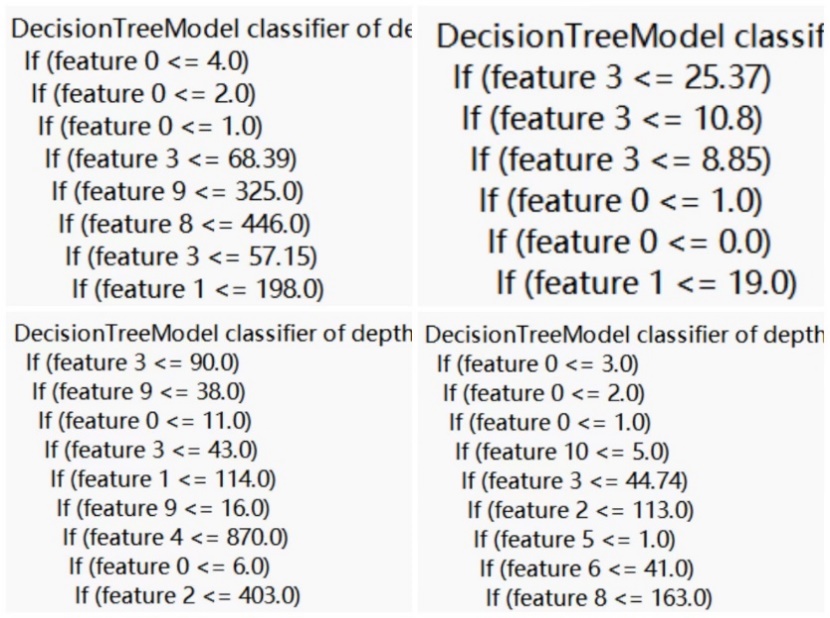


图2-37 决策树第一轮判断：上贝壳网，下安居客；左买房，右租房

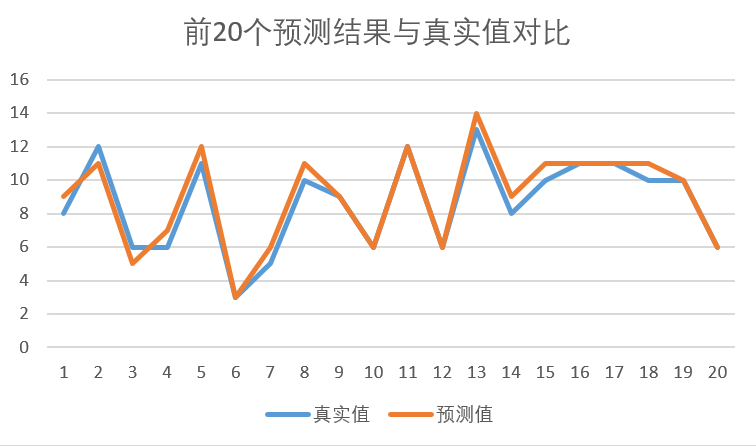


图2-38 前20个预测结果与真实值对比

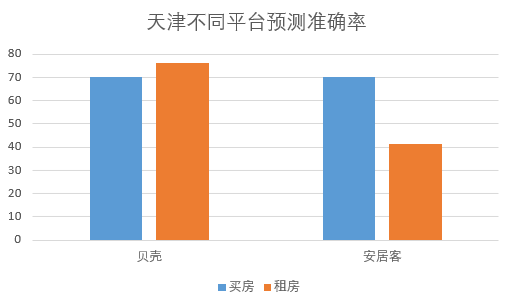


图2-39 天津不同平台预测准确率

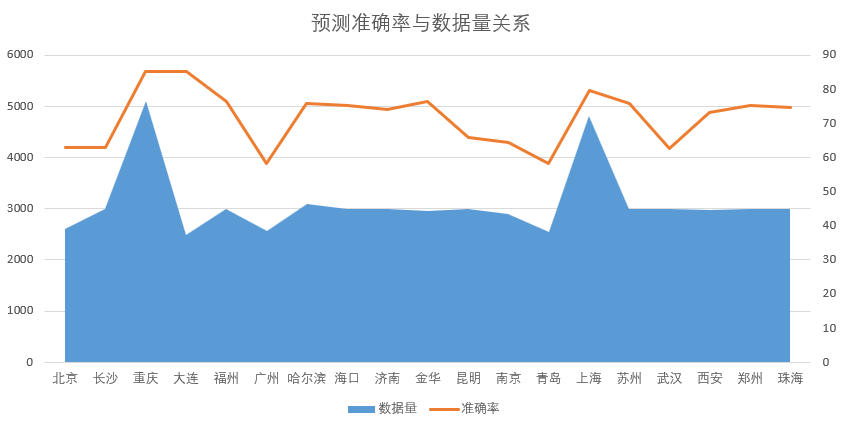


图2-40 预测准确率与数据量关系

通过对决策树第一轮判断的分析，我们发现城市区名（feature0）和面积（feature3）是被优先判断的两个最重要的特征，随之判断层数（feature9）和关键词（feature10）。而在区名中尤其是所在区的影响最为显著。这强调了地段和面积对于房价的重要性。

对比预测结果和真实值，我们发现模型整体上表现良好，训练数据量严重影响决策树模型性能。在实际应用中，应尽可能提供充足的训练数据以获得更好的预测性能，在数据量少的情况下可以考虑提升训练数据的比例。

# 总结

## 3.1 整体总结

在本次实验对于清洗后的数据，分别运用了MapReduce生成词云、K-means聚类和决策树预测方法。MapReduce生成词云的优势在于对大规模数据集的高效精炼，以及展示交易平台最频繁关键词的直观性；劣势在于非常有限的信息量与上下文信息的丢失。K-means的势在于可以帮助识别市场上的不同细分区域，发掘潜在的类别和关联；劣势在于此算法每簇的大小和形状较为理想，不一定符合现实中的分类情况，而且聚类的结果需要相关知识人为解释。决策树预测的优势在于树状结构显示不同特征对价格影响的清晰性以及对非线性关系的拟合能力，劣势在于可能存在过拟合问题。

## 3.2 反思与展望

本次实验在挖掘的深度上尚有欠缺，未来计划进一步扩展数据爬取范围（如评论数据、公开论坛数据等），以及进一步运用回归方法探索房屋特征的内在联系。

# 附录

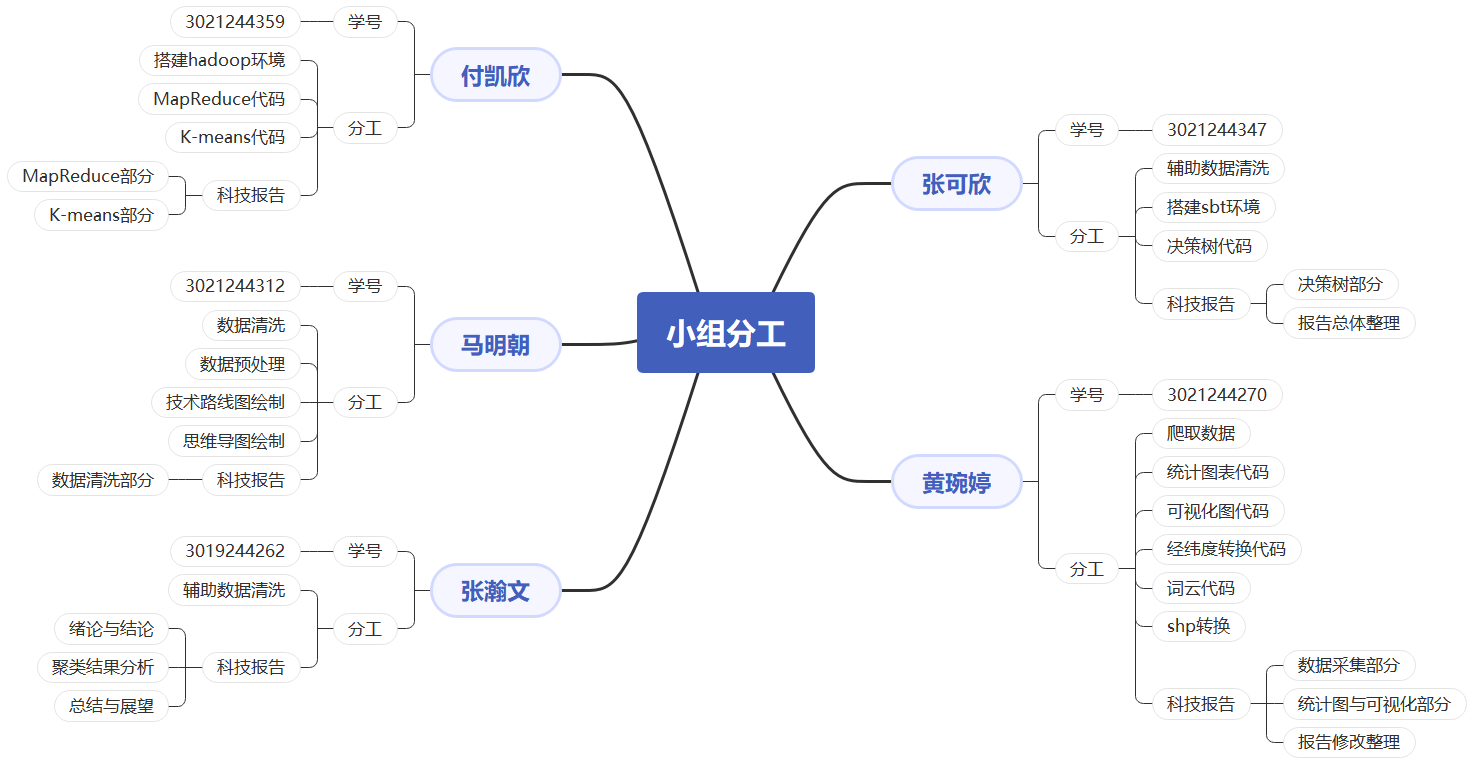


图1 小组分工图

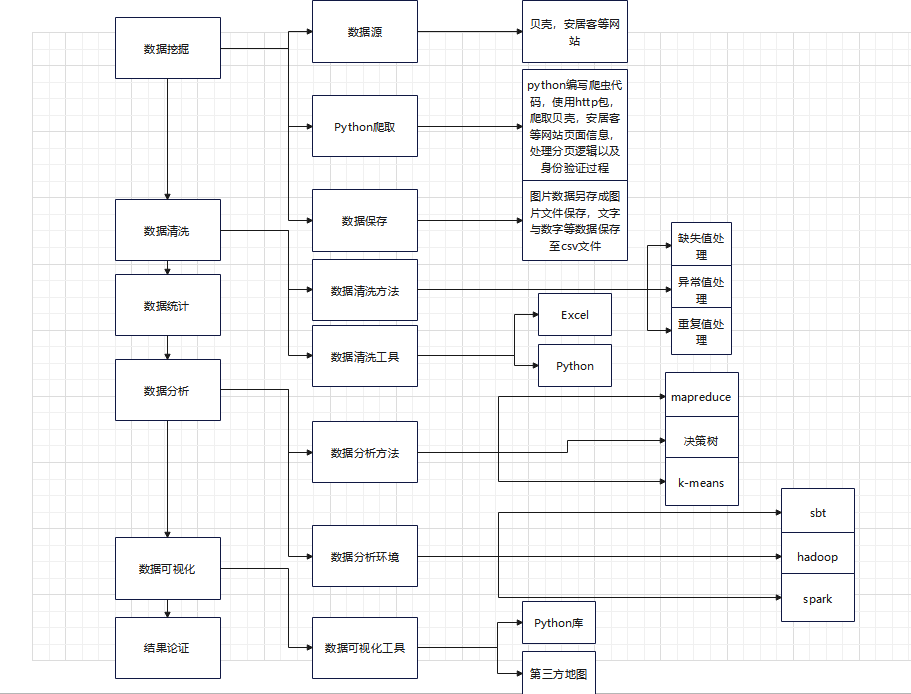


图2 技术路线图