# 房屋租赁数据处理与分析

**一、题目简介**

本题目是根据所学的大数据基础上的相关知识对提供的北京地区的房屋租赁数据进行分析，并以可视化的形势实现分析结果的可视化。包括价格分布、各区域租域、租赁类型、地理、时间分布等。

**二、数据集说明**

本次使用的数据集是2019 年 4 月 17 日公开的北京地区数据集Beijing, China，这里选择了其中的短租房源基础信息表listings.csv，包括房源、房东、位置、类型、价格、评论数量和可租时间等等。数据集为csv格式，共包含25211项数据，每项数据包含以下字段：

| **字段名** | **含义** | **示例** |
| --- | --- | --- |
| id | 房间编号 | 整型，"9567683" |
| name | 房间名称 | 字符串，"【青蕖】朝阳北路/大悦城/十里堡/6号线地铁，一室一厅独享整套阳光，可月租" |
| host\_id | 房东编号 | 整型，"43175532" |
| host\_name | 房东名称 | 字符串，"孜孜" |
| neighbourhood\_group | 所属区域组，在本城市中无实际含义。 | 字符串，无。 |
| neighbourhood | 行政区域 | 字符串，"朝阳区 / Chaoyang" |
| latitude | 纬度 | 浮点型，"39.92218" |
| longitude | 经度 | 浮点型，"116.50494" |
| room\_type | 房间类型（整套、独立房间、床位） | 字符串，"Entire home/apt" |
| price | 价格 | 整型，"343" |
| minimum\_nights | 最少住几晚 | 整型，"1" |
| number\_of\_reviews | 评论数 | 整型，"166" |
| last\_review | 上一次评论的时间 | 时间戳，"2021-02-08" |
| reviews\_per\_month | 平均每月评论数 | 浮点数，"2.60" |
| calculated\_host\_listings\_count | 房间数 | 整型，"2" |
| availability\_365 | 一年内可租用天数 | 整型，"295" |

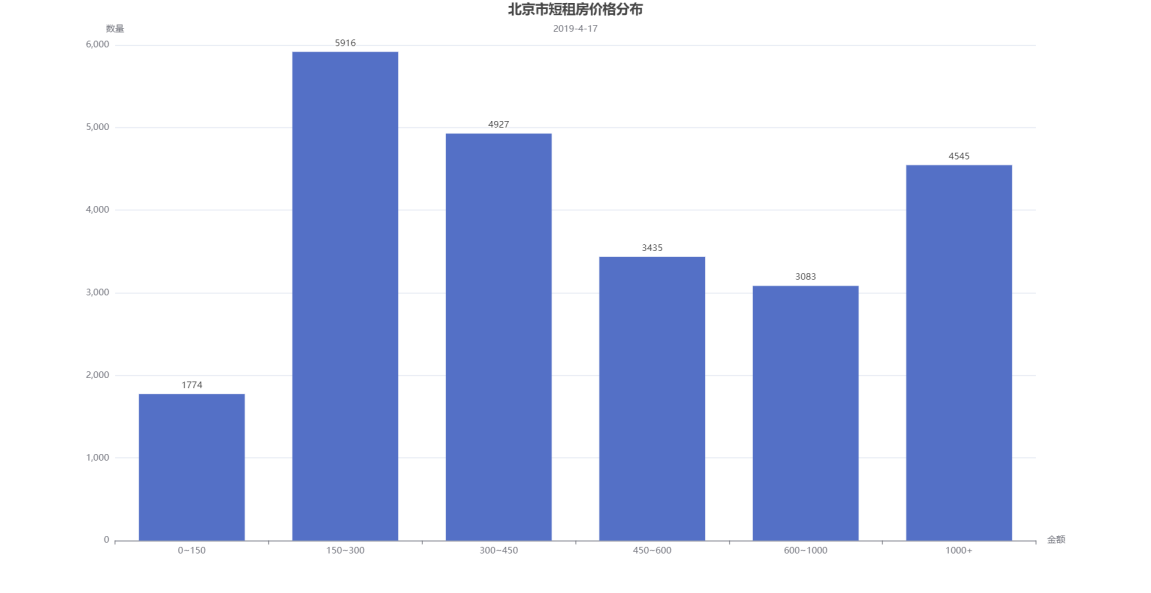
**三、步骤概述**

完成课题，大致包含以下5个步骤：  
（1）建立环境。（Hadop、HBase、Hive、MySQL、Sqoop)  
（2）创建数据源读取数据,并将数据存储进MySQL。  
（3）对数据集指定转换操作，保存到HBase。

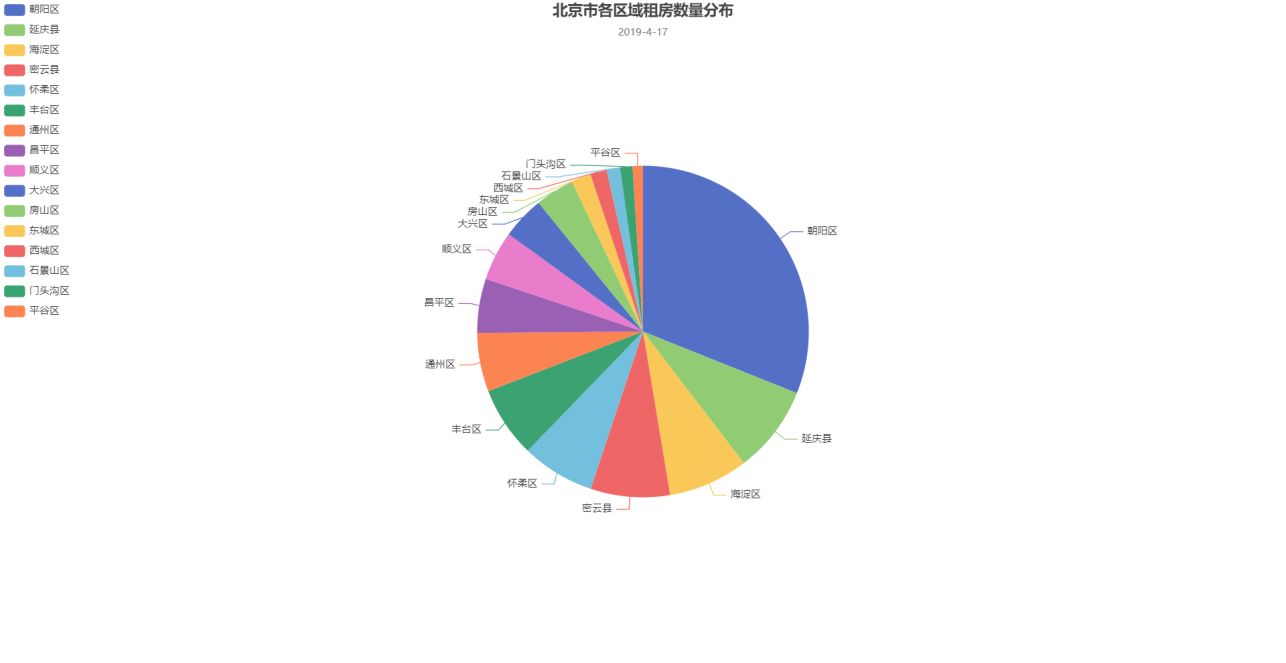
（4）利用Hive实数据进行分析，并将结果保存到MySQL。  
（5）对结果进行可视化。

**四、分析结果样例**

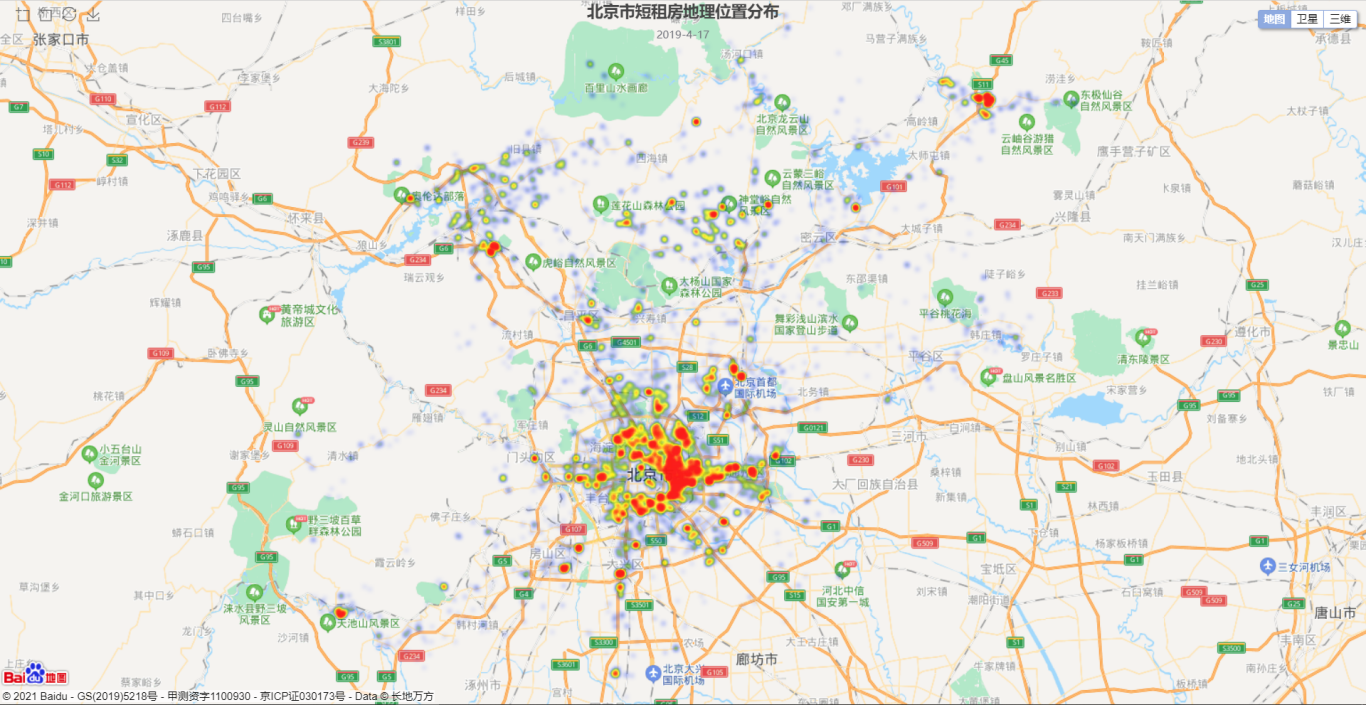
#### （1）价格分布



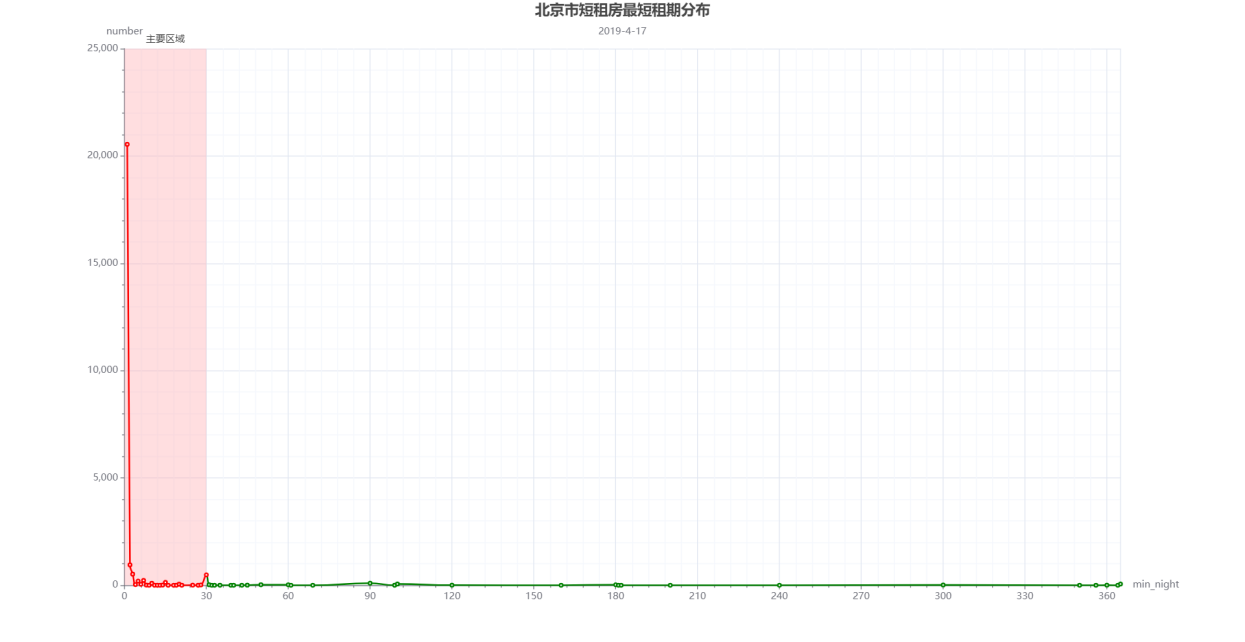
#### （2）各区域租房数量分布



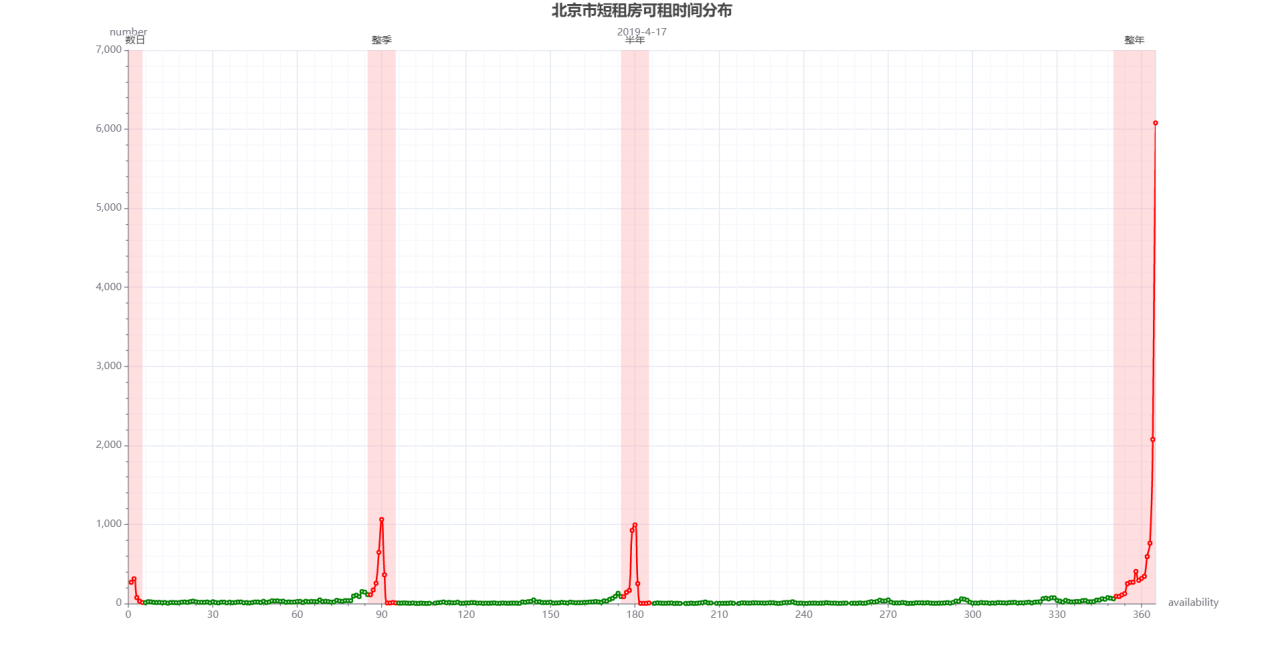
#### （3）地理位置分布

如图所示，大部分房源都位于朝阳区，与上面各区域租房数量相对应。  


#### （4）最短租期分布

如下图所示：绝大部分短租房都支持最短租期为1天。  


#### （5）可租时间分布



#### （6）房屋类型分布IMG_261

#### （7）各区域房价分布

